UNIT-1
FREQUENCY DISTRIBUTION

Structure:

1.0 Introduction
1.1 Objectives
1.2 Measures of Central Tendency
   1.2.1 Arithmetic mean
   1.2.2 Median
   1.2.3 Mode
   1.2.4 Empirical relation among mode, median and mode
   1.2.5 Geometric mean
   1.2.6 Harmonic mean
1.3 Partition values
   1.3.1 Quartiles
   1.3.2 Deciles
   1.3.3 Percentiles
1.4 Measures of dispersion
   1.4.1 Range
   1.4.2 Semi-interquartile range
   1.4.3 Mean deviation
   1.4.4 Standard deviation
   1.4.5 Geometric mean
1.5 Absolute and relative measure of dispersion
1.6 Moments
1.7 Karl Pearson’s β and γ coefficients
1.8 Skewness
1.9 Kurtosis
1.10 Let us sum up
1.11 Check your progress: The key.
1.0 INTRODUCTION

According to Simpson and Kafka a measure of central tendency is typical value around which other figures aggregate’.

According to Croxton and Cowden ‘An average is a single value within the range of the data that is used to represent all the values in the series. Since an average is somewhere within the range of data, it is sometimes called a measure of central value’.

1.1 OBJECTIVES

The main aim of this unit is to study the frequency distribution. After going through this unit you should be able to:

- describe measures of central tendency;
- calculate mean, mode, median, G.M., H.M.;
- find out partition values like quartiles, deciles, percentiles etc;
- know about measures of dispersion like range, semi-inter-quartile range, mean deviation, standard deviation;
- calculate moments, Karls Pearson’s β and γ coefficients, skewness, kurtosis.

1.2 MEASURES OF CENTRAL TENDENCY

The following are the five measures of average or central tendency that are in common use:

(i) Arithmetic average or arithmetic mean or simple mean
(ii) Median
(iii) Mode
(iv) Geometric mean
(v) Harmonic mean

Arithmetic mean, Geometric mean and Harmonic means are usually called Mathematical averages while Mode and Median are called Positional averages.
1.2.1 ARITHMETIC MEAN

To find the arithmetic mean, add the values of all terms and them divide sum by the number of terms, the quotient is the arithmetic mean. There are three methods to find the mean:

(i) Direct method: In individual series of observations \( x_1, x_2, ..., x_n \) the arithmetic mean is obtained by following formula.
\[
\text{A.M.} = \frac{x_1 + x_2 + x_3 + x_4 + \ldots + x_{n-1} + x_n}{n}
\]

(ii) Short-cut method: This method is used to make the calculations simpler.
Let \( A \) be any assumed mean (or any assumed number), \( d \) the deviation of the arithmetic mean, then we have
\[
M = A + \frac{\sum fd}{N} \quad (\because \quad d=(x-A))
\]

(iii) Step deviation method: If in a frequency table the class intervals have equal width, say \( i \) than it is convenient to use the following formula.
\[
M = A + \frac{\sum fui}{n} \times i
\]
where \( u=(x-A)/i \), and \( i \) is length of the interval, \( A \) is the assumed mean.

Example 1. Compute the arithmetic mean of the following by direct and short-cut methods both:

<table>
<thead>
<tr>
<th>Class</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-30</td>
<td>8</td>
</tr>
<tr>
<td>30-40</td>
<td>26</td>
</tr>
<tr>
<td>40-50</td>
<td>30</td>
</tr>
<tr>
<td>50-60</td>
<td>20</td>
</tr>
<tr>
<td>60-70</td>
<td>16</td>
</tr>
</tbody>
</table>

Solution.

<table>
<thead>
<tr>
<th>Class</th>
<th>Mid Value</th>
<th>( f )</th>
<th>( fx )</th>
<th>( d = x-A )</th>
<th>( f \cdot d )</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-30</td>
<td>25</td>
<td>8</td>
<td>200</td>
<td>-20</td>
<td>-160</td>
</tr>
<tr>
<td>30-40</td>
<td>35</td>
<td>26</td>
<td>910</td>
<td>-10</td>
<td>-260</td>
</tr>
<tr>
<td>40-50</td>
<td>45</td>
<td>30</td>
<td>1350</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>50-60</td>
<td>55</td>
<td>20</td>
<td>1100</td>
<td>10</td>
<td>200</td>
</tr>
<tr>
<td>60-70</td>
<td>65</td>
<td>16</td>
<td>1040</td>
<td>20</td>
<td>320</td>
</tr>
</tbody>
</table>

By direct method
\[
M = \frac{\sum fx}{N} = \frac{4600}{100} = 46.
\]

By short cut method.

Let assumed mean \( A = 45 \).
\[ M = A + \frac{(\sum fd)}{N} = 45 + \frac{100}{100} = 46. \]

Example 2: Compute the mean of the following frequency distribution using step deviation method:

<table>
<thead>
<tr>
<th>Class</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-11</td>
<td>9</td>
</tr>
<tr>
<td>11-22</td>
<td>17</td>
</tr>
<tr>
<td>22-33</td>
<td>28</td>
</tr>
<tr>
<td>33-44</td>
<td>26</td>
</tr>
<tr>
<td>44-55</td>
<td>15</td>
</tr>
<tr>
<td>55-66</td>
<td>8</td>
</tr>
</tbody>
</table>

Solution.

<table>
<thead>
<tr>
<th>Class</th>
<th>Mid-Value</th>
<th>f</th>
<th>(d = x - A) ((A = 38.5))</th>
<th>(u = (x - A)/i)</th>
<th>(fu)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-11</td>
<td>5.5</td>
<td>9</td>
<td>-33</td>
<td>-3</td>
<td>-27</td>
</tr>
<tr>
<td>11-22</td>
<td>16.5</td>
<td>17</td>
<td>-22</td>
<td>-2</td>
<td>-34</td>
</tr>
<tr>
<td>22-33</td>
<td>27.5</td>
<td>28</td>
<td>-11</td>
<td>-1</td>
<td>-28</td>
</tr>
<tr>
<td>33-44</td>
<td>38.5</td>
<td>26</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>44-55</td>
<td>49.5</td>
<td>15</td>
<td>11</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>55-66</td>
<td>60.5</td>
<td>8</td>
<td>22</td>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>N = 103</td>
<td></td>
<td>(\sum fu = -58)</td>
<td></td>
</tr>
</tbody>
</table>

Let the assumed mean \(A = 38.5\), then

\[ M = A + i(\sum fu)/N = 38.5 + 11(-58)/103 \]

\[ = 38.5 - \frac{638}{103} = 38.5 - 6.194 = 32.306 \]

**PROPERTIES OF ARITHMETIC MEAN**

Property 1: The algebraic sum of the deviations of all the variates from their arithmetic mean is zero.

**Proof.** Let \(X_1, X_2, \ldots, X_n\) be the values of the variates and their corresponding frequencies be \(f_1, f_2, \ldots, f_n\) respectively.

Let \(x_i\) be the deviation of the variate \(X_i\) from the mean \(M\), where \(i = 1, 2, \ldots, n\). Then

\[ x_i = X_i - M, \quad i = 1, 2, \ldots, n. \]

\[ \sum_{i=1}^{n} f_i x_i = \sum_{i=1}^{n} f_i (X_i - M) \]

\[ = M \sum_{i=1}^{n} f_i - M \sum_{i=1}^{n} f_i \]
Exercise 1(a)

Q.1) Marks obtained by 9 students in statistics are given below.

52  75  40  70  43  65  40  35  48

calculate the arithmetic mean.

Q.2) Calculate the arithmetic mean of the following distribution

Variate :  6  7  8  9  10  11  12
Frequency: 20  43  57  61  72  45  39

Q.3) Find the mean of the following distribution

Variate :  0-10  10-20  20-30  30-40  40-50
Frequency: 31  44  39  58  12

1.2.2 MEDIAN

The median is defined as the measure of the central term, when the given terms (i.e., values of the variate) are arranged in the ascending or descending order of magnitudes. In other words the median is value of the variate for which total of the frequencies above this value is equal to the total of the frequencies below this value.

Due to Corner, “The median is the value of the variable which divides the group into two equal parts one part comprising all values greater, and the other all values less then the median”.

For example. The marks obtained, by seven students in a paper of Statistics are 15, 20, 23, 32, 34, 39, 48 the maximum marks being 50, then the median is 32 since it is the value of the 4th term, which is situated such that the marks of 1st, 2nd and 3rd students are less than this value and those of 5th, 6th and 7th students are greater then this value.

COMPUTATION OF MEDIAN

(a)Median in individual series.

Let \( n \) be the number of values of a variate (i.e. total of all frequencies). First of all we write the values of the variate (i.e., the terms) in ascending or descending order of magnitudes

Here two cases arise:
Case 1. If \( n \) is odd then value of \( (n+1)/2 \) th term gives the median.

Case 2. If \( n \) is even then there are two central terms i.e., \( n/2 \) th and \( (n+1)/2 \) th term. The mean of these two values gives the median.

(b) Median in continuous series (or grouped series). In this case, the median \( (M_d) \) is computed by the following formula

\[
M_d = l + \frac{n}{2} - cf \times \frac{f}{i}
\]

Where

- \( M_d = \) median
- \( l = \) lower limit of median class
- \( cf = \) total of all frequencies before median class
- \( f = \) frequency of median class
- \( i = \) class width of median class.

Example 1 – According to the census of 1991, following are the population figure, in thousands, of 10 cities:

1400, 1250, 1670, 1800, 700, 650, 570, 488, 2100, 1700.

Find the median.

Solution. Arranging the terms in ascending order.

488, 570, 650, 700, 1250, 1400, 1670, 1800, 2100.

Here \( n=10 \), therefore the median is the mean of the measure of the 5 th and 6 th terms.

Here 5 th term is 1250 and 6 th term is 1400.

Median \( (Md) = (1250+14000)/2 \) Thousands

\[ = 1325 \text{ Thousands} \]

Examples 2. Find the median for the following distribution:

<table>
<thead>
<tr>
<th>Wages in Rs.</th>
<th>0-10</th>
<th>10-20</th>
<th>20-30</th>
<th>30-40</th>
<th>40-50</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of workers</td>
<td>22</td>
<td>38</td>
<td>46</td>
<td>35</td>
<td>20</td>
</tr>
</tbody>
</table>

Solution. We shall calculate the cumulative frequencies.
Here \( N = 161 \). Therefore median is the measure of \( (N + 1)/2 \)\(^{th}\) term i.e \( 81^{st} \) term. Clearly \( 81^{st} \) term is situated in the class 20-30. Thus 20-30 is the median class. Consequently.

\[
\text{Median } M_d = l + \left( \frac{n - \text{cf}}{f} \right) \times i
\]

\[
= 20 + (\frac{1}{2} \times 161 - 60) / 46 \times 10
\]

\[
= 20 + 205/46 = 20 + 4.46 = 24.46.
\]

Example 3. Find the median of the following frequency distribution:

<table>
<thead>
<tr>
<th>Marks</th>
<th>No. of students</th>
<th>Marks</th>
<th>No. of students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 10</td>
<td>15</td>
<td>Less than 50</td>
<td>106</td>
</tr>
<tr>
<td>Less than 20</td>
<td>35</td>
<td>Less than 60</td>
<td>120</td>
</tr>
<tr>
<td>Less than 30</td>
<td>60</td>
<td>Less than 70</td>
<td>125</td>
</tr>
<tr>
<td>Less than 40</td>
<td>84</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Solution. The cumulative frequency distribution table:

<table>
<thead>
<tr>
<th>Class (Marks)</th>
<th>Frequency f (No. of students)</th>
<th>Cumulative Frequency (C. F.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-10</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>10-20</td>
<td>20</td>
<td>35</td>
</tr>
<tr>
<td>20-30</td>
<td>25</td>
<td>60</td>
</tr>
<tr>
<td>30-40</td>
<td>24</td>
<td>84</td>
</tr>
<tr>
<td>40-50</td>
<td>22</td>
<td>106</td>
</tr>
<tr>
<td>50-60</td>
<td>14</td>
<td>120</td>
</tr>
<tr>
<td>60-70</td>
<td>5</td>
<td>125</td>
</tr>
</tbody>
</table>

Total \( N = 125 \)

\[
\text{Median} = \text{measure of } \left( \frac{125 + 1}{2} \right)^{th} \text{ term}
\]
Clearly 63rd term is situated in the class 30-40.

Thus median class  = 30 - 40

\[
\text{Median } M_d = l + \frac{n - cf}{f} \times i
\]

\[
= 30 + \frac{(125/2 - 60)}{24} \times 10
\]

\[
= 30 + 25/24
\]

\[
= 30 + 1.04 = 31.04
\]

1.2.3 MODE

The word ‘mode is formed from the French word ‘La mode’ which means ‘in fashion’. According to Dr. A. L. Bowle ‘the value of the graded quantity in a statistical group at which the numbers registered are most numerous, is called the mode or the position of greatest density or the predominant value.’

1.2.3 MODE

According to other statisticians, ‘The value of the variable which occurs most frequently in the distribution is called the mode.’

“The mode of a distribution is the value around the items tends to be most heavily concentrated. It may be regarded at the most typical value of the series”.

Definition. The mode is that value (or size) of the variate for which the frequency is maximum or the point of maximum frequency or the point of maximum density. In other words, the mode is the maximum ordinate of the ideal curve which gives the closest fit to the actual distribution.
Method to Compute the mode:
(a) When the values (or measures) of all the terms (or items) are given. In this case the mode is the value (or size) of the term (or item) which occurs most frequently.

Example 1. Find the mode from the following size of shoes

<table>
<thead>
<tr>
<th>Size of shoes</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Here maximum frequency is 3 whose term value is 6. Hence the mode is modal size number 6.

(b) In continuous frequency distribution the computation of mode is done by the following formula

\[
\text{Mode } M_0 = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i
\]

\[ l = \text{lower limit of class,} \]
\[ f_1 = \text{frequency of modal class,} \]
\[ f_0 = \text{frequency of the class just preceding to the modal class,} \]
\[ f_2 = \text{frequency of the class just following of the modal class,} \]
\[ i = \text{class interval} \]

Example 2. Compute the mode of the following distribution:

<table>
<thead>
<tr>
<th>Class</th>
<th>0-7</th>
<th>7-14</th>
<th>14-21</th>
<th>21-28</th>
<th>28-35</th>
<th>35-42</th>
<th>42-49</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>19</td>
<td>25</td>
<td>36</td>
<td>72</td>
<td>51</td>
<td>43</td>
<td>28</td>
</tr>
</tbody>
</table>

Solution. Here maximum frequency 72 lies in the class-interval 21-28. Therefore 21-28 is the modal class.

\[ l = 21, \quad f_1 = 72, \quad f_0 = 36, \quad f_2 = 51, \quad i = 7 \]

\[
\text{Mode } M_0 = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i
\]

\[ = 21 + \frac{72 - 36}{(2 \times 72 - 36 - 51)} \times 10 \]
\[ = 21 + \frac{357}{87} \]
\[ = 21 + 4.103 \]
\[ = 25.103. \]
(c) Method of determining mode by the method of grouping frequencies. This method is usually applied in the cases when there are two maximum frequencies against two different size of items. This method is also applied in the cases when it is possible that the effect of neighboring frequencies on the size of item (of maximum frequency) may be greater. The method is as follows:

Firstly the items are arranged in ascending or descending order and corresponding frequencies are written against them. The frequencies are then grouped in two and then in threes and then is fours (if necessary). In the first stage of grouping, they are grouped (i.e., frequencies are added) by taking, first and second, third and fourth, ..., . After it, the frequencies are added in threes. The frequencies are added in the following two ways:

1. (i) First and second, third and fourth, fifth and sixth, seventh and eighth, ...
   (ii) Second and third, fourth and fifth, ...
2. (i) First, second and third; fourth, fifth and sixth, ...
   (ii) Second, third and fourth; fifth, sixth and seventh, ...
   (iii) Third, fourth and fifth; sixth seventh and eighth, ...

Now the items with maximum frequencies are selected and the item which contains the maximum is called the mode. For illustration see following example 1.

Example 3. Compute the mode from the following distribution:

<table>
<thead>
<tr>
<th>Size of Item</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>9</td>
<td>12</td>
<td>14</td>
<td>14</td>
<td>15</td>
<td>11</td>
<td>13</td>
</tr>
</tbody>
</table>

Solution. From the given data we observe that size 11 has the maximum frequency 15, but it is possible that the effect of neighboring frequencies on the size of the item may be greater. Thus it may happen that the frequencies of size 10 or 12 may be greater and 11 may not remain mode. We shall apply the method of grouping.

<table>
<thead>
<tr>
<th>Size of Items</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2</td>
<td>7</td>
<td>13</td>
<td>15</td>
<td>22</td>
<td>29</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>17</td>
<td>21</td>
<td>22</td>
<td>22</td>
<td>29</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>17</td>
<td>21</td>
<td>22</td>
<td>22</td>
<td>29</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
<td>21</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>8</td>
<td>12</td>
<td>26</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>9</td>
<td>14</td>
<td>26</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
</tr>
</tbody>
</table>
We have used brackets against the frequencies which have been grouped. Now we shall find the size of the item containing maximum frequency:

<table>
<thead>
<tr>
<th>Column</th>
<th>Size of item having maximum frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>11</td>
</tr>
<tr>
<td>II</td>
<td>10,11</td>
</tr>
<tr>
<td>III</td>
<td>9,10</td>
</tr>
<tr>
<td>IV</td>
<td>10,11,12</td>
</tr>
<tr>
<td>V</td>
<td>8,9,10</td>
</tr>
<tr>
<td>VI</td>
<td>9,10,11</td>
</tr>
</tbody>
</table>

Here size 8 occurs 1 time, 9 occurs 3 times, 10 occurs 5 times, 11 occurs 4 times, 12 occurs 1 time.
Since 10 occurs maximum number of times (5 times).
Hence the required mode is size 10.

1.2.4 EMPIRICAL RELATION BETWEEN MEDIAN AND MODE

For moderately asymmetrical distribution (or for asymmetrical curve), the relation
Mean – Mode = 3 (Mean - Median),
approximately holds. In such a case, first evaluate mean and median and then mode is determined by
Mode = 3 Median – 2 Mean.
If in the asymmetrical curve the area on the left of mode is greater than area on the right then
Mean < median < mode, i.e., (M < Md < M0)
If in the asymmetrical curve, the area on the left of mode is less than the area on the right then in this case

Mode < median , mean, i.e. (M0 < Md < M).

Exercise 1(c)

Q.1) Find the Mode of the following model size number of shoes.
Model size no. of shoes : 3,4,2,1,7,6,7,5,6,8,9,5.

Q.2) Compute the Mode of the following distribution.
Class                : 0-7  7-14  14-21  21-28  28-35  35-42  42-49
Frequency            : 19  25  36  72  51  43  28

1.2.5 GEOMETRIC MEAN

If x₁,x₂, ..., xₙ are n values of the variate x, none of which is zero . Then their geometric mean G is defined by

G = (x₁, x₂, ..., xₙ)¹/n

If f₁, f₂, ..., fₙ are the frequencies of x₁,x₂,..., xₙ respectively, then geometric mean G is given by

G = {x₁ f₁ x₂ f₂ ... xₙ fₙ}¹/N
N = f₁ + f₂ + .... + fₙ

Taking log of (1), we get

Log G = 1/N [f₁ log x₁ + f₂ log x₂ + ... + fₙ log xₙ]

Log G = \( \frac{1}{N} \sum_{i=1}^{n} f_i \log x_i \)

Log G = \( \frac{1}{N} \sum_{i=1}^{n} f \log x \).

Example 2. Compute the geometric mean of the following distribution:
# Marks

<table>
<thead>
<tr>
<th>Marks</th>
<th>No. of students</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-10</td>
<td>5</td>
</tr>
<tr>
<td>10-20</td>
<td>8</td>
</tr>
<tr>
<td>20-30</td>
<td>3</td>
</tr>
<tr>
<td>30-40</td>
<td>4</td>
</tr>
</tbody>
</table>

## Solution

<table>
<thead>
<tr>
<th>Class</th>
<th>Mid-value</th>
<th>Frequency</th>
<th>Log ( \log_{10} x )</th>
<th>Product ( F \log x )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-10</td>
<td>5</td>
<td>5</td>
<td>0.6990</td>
<td>3.4950</td>
</tr>
<tr>
<td>10-20</td>
<td>15</td>
<td>8</td>
<td>1.1761</td>
<td>9.4088</td>
</tr>
<tr>
<td>20-30</td>
<td>25</td>
<td>3</td>
<td>1.3979</td>
<td>4.1937</td>
</tr>
<tr>
<td>30-40</td>
<td>35</td>
<td>4</td>
<td>1.5441</td>
<td>6.1764</td>
</tr>
</tbody>
</table>

\[ N = \sum f = 20 \]

\[ \sum f \log x = 23.2739 \]

\[ \log G = \frac{\sum \log x}{N} = \frac{23.2739}{20} = 1.1637 \]

\[ G = \text{anti-log} (1.1637) = 12.58 \text{ marks}. \]

## 1.2.6 HARMONIC MEAN

The Harmonic mean of a series of values is the reciprocal of the arithmetic means of their reciprocals. Thus if \( x_1, x_2, \ldots, x_n \) (none of them being zero) is a series and \( H \) is its harmonic mean then

\[
\frac{1}{H} = \frac{1}{N} \left[ \frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n} \right]
\]

If \( f_1, f_2, \ldots, f_n \) be the frequencies of \( x_1, x_2, \ldots, x_n \) (none of them being zero) then harmonic mean \( H \) is given by

\[
H.M. = \frac{\sum f}{\sum f \times \frac{1}{x}}
\]

### Example 1.

Find the harmonic mean of the marks obtained in a class test, given below

<table>
<thead>
<tr>
<th>Marks</th>
<th>No. of students</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>13</td>
<td>8</td>
</tr>
<tr>
<td>14</td>
<td>5</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
</tr>
</tbody>
</table>

## Solution

<table>
<thead>
<tr>
<th>Marks ( X )</th>
<th>Frequency ( f )</th>
<th>( 1/x )</th>
<th>( f \times 1/x )</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>3</td>
<td>0.0909</td>
<td>0.2727</td>
</tr>
<tr>
<td>12</td>
<td>7</td>
<td>0.0833</td>
<td>0.5831</td>
</tr>
<tr>
<td>13</td>
<td>8</td>
<td>0.0769</td>
<td>0.6152</td>
</tr>
<tr>
<td>14</td>
<td>5</td>
<td>0.0714</td>
<td>0.3570</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>0.0667</td>
<td>0.1334</td>
</tr>
</tbody>
</table>
Required harmonic mean is given by

\[ H.M. = \frac{\sum f}{\sum f \times \frac{1}{x}} \]

\[ = \frac{25}{1.9614} \]

\[ = \frac{25}{1.9614} \]

\[ = \frac{250000}{19614} \]

\[ = 12.746 \text{ marks.} \]

Property. For two observations \( x_1 \) and \( x_2 \), we have

\[ AH = G^2 \]

Where \( A = \) arithmetic mean, \( H = \) harmonic mean and \( G = \) geometric mean.

1.3. PARTITION VALUES

If the values of the variate are arranged in ascending or descending order of magnitudes then we have seen above that median is that value of the variate which divides the total frequencies in two equal parts. Similarly the given series can be divided into four, ten and hundred equal parts. The values of the variate dividing into four equal parts are called Quartile, into ten equal parts are called Decile and into hundred equal parts are called Percentile.

1.3.1 QUARTILES :

Definition. The values of the variate which divide the total frequency into four equal parts, are called quartiles. That value of the variate which divides the total frequency into two equal parts is called median. The lower quartile or first quartile denoted by \( Q_1 \) divides the frequency between the lowest value and the median into two equal parts and similarly the upper quartile (or third quartile) denoted by \( Q_3 \) divides the frequency between the median and the greatest value into two equal parts. The formulas for computation of quartiles are given by

\[ Q_1 = l + \frac{n - cf}{f} \times i \]

\[ Q_3 = l + \frac{3n - cf}{f} \times i \]

where \( l, cf, n, f, i \) have the same meaning as in the formula for median.

1.3.2 DECILES :
Definition. The values of the variate which divide the total frequency into ten equal parts are called deciles. The formulas for computation are given by

\[ D_1 = l + \frac{n - cf}{f} \times i, \quad D_2 = l + \frac{2n - cf}{f} \times i \text{ etc...} \]

1.3.3 PERCENTILES:

Definition. The values of the variate which divide the total frequency into hundred equal parts, are called percentiles. The formulas for computation are:

\[ P_1 = l + \frac{n - cf}{f} \times i, \quad P_{70} = l + \frac{70n - cf}{f} \times i \text{ etc...} \]

Example 1. Compute the lower and upper quartiles, fourth decile and 70\textsuperscript{th} percentile for the following distribution:

<table>
<thead>
<tr>
<th>Marks group</th>
<th>No. of students</th>
<th>Marks group</th>
<th>No. of students</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-10</td>
<td>5</td>
<td>25-30</td>
<td>5</td>
</tr>
<tr>
<td>10-15</td>
<td>6</td>
<td>30-35</td>
<td>4</td>
</tr>
<tr>
<td>15-20</td>
<td>15</td>
<td>35-40</td>
<td>2</td>
</tr>
<tr>
<td>20-25</td>
<td>10</td>
<td>40-45</td>
<td>2</td>
</tr>
</tbody>
</table>

Solution. First we make the cumulative frequency table:

<table>
<thead>
<tr>
<th>Class</th>
<th>Frequency</th>
<th>Cumulative Frequency</th>
<th>Class</th>
<th>Frequency</th>
<th>Cumulative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-10</td>
<td>5</td>
<td>5</td>
<td>25-30</td>
<td>5</td>
<td>41</td>
</tr>
<tr>
<td>10-15</td>
<td>6</td>
<td>11</td>
<td>30-35</td>
<td>4</td>
<td>45</td>
</tr>
<tr>
<td>15-20</td>
<td>15</td>
<td>26</td>
<td>35-40</td>
<td>2</td>
<td>47</td>
</tr>
<tr>
<td>20-25</td>
<td>10</td>
<td>36</td>
<td>40-45</td>
<td>2</td>
<td>49</td>
</tr>
</tbody>
</table>

(i) To compute Q\textsubscript{1}. Here N = 49, \( \frac{1}{4} N = \frac{1}{4} \times 9 = 12.25 \) which clearly lies in 15-20
Thus 15-20 is lower quartile class.

\[ l = 15, \ cf = 11, \ f = 15, \ i = 20-15 = 5 \]

\[ Q_1 = l + \frac{n - cf}{f} \times i \]

\[ = 15 + (12.25 - 11) / 15 \times 5 = 15 + 0.417 = 15.417. \]

(ii) To Compute Q\textsubscript{3}. Here \( \frac{3}{4} N = \frac{3}{4} \times 49 = 36.75 \) which clearly lies in the class 25-30.

Thus \( l = 25, \ cf = 36, \ f = 5, \ i = 30-25 = 5 \)
\[ Q_3 = l + \frac{3n - cf}{f} \times i \]

\[ = 25 + (36.75 - 36)/5 \times 5 = 25 + 0.75 = 25.75 \]

(iii) To compute \( D_4 \) Here \( 4/10 \ N = 4/10 \times 49 = 19.6 \), which clearly lies in the class 15-20.

Thus \( l = 15, \ cf = 11, \ f = 15, \ i = 5 \)

\[ D_4 = l + \frac{4n - cf}{f} \times i \]

\[ = 15 + 19.6 - 11/15 \times 5 = 15 + 2.87 = 17.87 \]

(iv) To compute \( P_{70} \) Here \( 70N/100 = 7/10 \times 49 = 34.3 \) which clearly lies in the class 20-25.

Thus \( l = 20, \ cf = 26, \ f = 10, \ i = 5 \)

\[ P_{70} = l + \frac{70n - cf}{f} \times i \]

\[ = 20 + (34.3 - 26)/10 \times 5 = 20 + 4.15 = 24.15 \]

---

**Exercise 1(b)**

**Q.1)** Find the median of the following.

| 20 | 18 | 22 | 27 | 25 | 12 | 15 |

**Q.2)** Calculate the median, lower and upper quartiles, third decile and 60\(^{th} \) percentile for the following distribution.

<table>
<thead>
<tr>
<th>Class</th>
<th>0-10</th>
<th>10-20</th>
<th>20-30</th>
<th>30-40</th>
<th>40-50</th>
<th>50-60</th>
<th>60-70</th>
<th>70-80</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>5</td>
<td>8</td>
<td>7</td>
<td>12</td>
<td>28</td>
<td>20</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

---

**1.4 MEASURES OF DISPERSION**

**DISPERSION OR VARIATION**

An averages gives an idea of central tendency of the given distribution but it is necessary to know how the variates are clustered around or scattered away from the average. To explain it more clearly consider the works of two typists who typed the following number of pages in 6 working days of a week:

<table>
<thead>
<tr>
<th>Mon.</th>
<th>Tues.</th>
<th>Wed.</th>
<th>Thurs.</th>
<th>Fri.</th>
<th>Sat.</th>
<th>Total pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 typist</td>
<td>15</td>
<td>20</td>
<td>25</td>
<td>25</td>
<td>30</td>
<td>35</td>
</tr>
</tbody>
</table>
We see that each of the typist 1 and 2 typed 150 pages in 6 working days and so the average in both the cases is 25. Thus there is no difference in the average, but we know that in the first case the number of pages varies from 15 to 35 while in the second case the number of pages varies from 10 to 40. This denotes that the greatest deviation from the mean in the first case is 10 and in the second case it is 15 i.e., there is a difference between the two series. The variation of this type is termed scatter or dispersion or spread.

Definition. The degree to which numerical data tend to spread about an average value is called variation or dispersion or spread of the data.

Various measures of dispersion or variation are available, the most common are the following.

1.4.1 THE RANGE

It is the simplest possible measure of dispersion. The range of a set of numbers (data) is the difference between the largest and the least numbers in the set i.e. values of the variable. If this difference is small then the series of numbers is supposed regular and if this difference is large then the series is supposed to be irregular.

Example: Compute the range for the following observation

| 15 | 20 | 25 | 25 | 30 | 35 |

Solution: Range = Largest – Smallest

i.e., 35-15=20

1.4.2 SEMI-INTER-QUARTILE RANGE

Definition. The inter quartile range of a set of data is defined by

\[ \text{Inter-quartile range} = Q_3 - Q_1 \]

where Q1 and Q3 are respectively the first and third quartiles for the data.

Semi-inter quartile range (or quartile deviation) is denoted by Q and is defined by

\[ Q = (Q_3 - Q_1)/2 \]

where Q1 and Q3 have the same meaning as given above.

The semi-inter-quartile range is a better measure of dispersion than the range and is easily computed. Its drawback is that it does not take into account all the items.
1.4.3 MEAN DEVIATION

Definition. The average (or mean) deviation about any point M, of a set of N numbers \( x_1, x_2, \ldots, x_N \) is defined by

\[
\text{Mean Deviation (M. D.)} = \delta_m = \frac{1}{N} \sum_{i=1}^{n} x_i - M
\]

where M is the mean or median or mode according as the mean deviation from the mean or median or mode is to be computed, \( l x_i - M l \) represents the absolute (or numerical) value. Thus \( l-5l = 5 \).

If \( x_1, x_2, \ldots, x_k \) occur with frequencies \( f_1, f_2, \ldots, f_k \) respectively, then the mean deviation (\( \delta_m \)) is defined by

\[
\delta_m = \frac{1}{N} \sum_{j=1}^{k} f_j(x_j - M) = \frac{1}{N} \sum f (x - M)
\]

Mean deviation depends on all the values of the variables and therefore it is a better measure of dispersion than the range or the quartile deviation. Since signs of the deviations are ignored (because all deviations are taken positive), some artificiality is created.

In case of grouped frequency distribution the mid-values are taken as \( x \).

Example 1. Find the mean deviation from the arithmetic mean of the following distribution:

<table>
<thead>
<tr>
<th>Marks</th>
<th>0-10</th>
<th>10-20</th>
<th>20-30</th>
<th>30-40</th>
<th>40-50</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of students</td>
<td>5</td>
<td>8</td>
<td>15</td>
<td>16</td>
<td>6</td>
</tr>
</tbody>
</table>

Solution. Let assumed mean \( A = 25 \) and \( i=10 \)

<table>
<thead>
<tr>
<th>Class</th>
<th>Mid value X</th>
<th>Frequency f</th>
<th>( u = \frac{x - A}{i} )</th>
<th>fu</th>
<th>x-M</th>
<th>f lx-Ml</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-10</td>
<td>5</td>
<td>5</td>
<td>-2</td>
<td>-10</td>
<td>-22</td>
<td>110</td>
</tr>
<tr>
<td>10-20</td>
<td>15</td>
<td>8</td>
<td>-1</td>
<td>-8</td>
<td>-12</td>
<td>96</td>
</tr>
<tr>
<td>20-30</td>
<td>25</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>30</td>
</tr>
<tr>
<td>30-40</td>
<td>35</td>
<td>16</td>
<td>1</td>
<td>16</td>
<td>8</td>
<td>128</td>
</tr>
<tr>
<td>40-50</td>
<td>45</td>
<td>6</td>
<td>2</td>
<td>12</td>
<td>18</td>
<td>108</td>
</tr>
<tr>
<td>Total</td>
<td>( \Sigma f = 50 )</td>
<td>( \Sigma fu = 10 )</td>
<td>( \Sigma f l x - M l )</td>
<td>472</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Arithmetic mean  \[ M = u + \frac{\sum f_i u}{N} \times i = 25 + 10/50 \times 10 = 27. \]

The required mean deviation from arithmetic mean

\[ \delta_m = \frac{\sum f(x - M)}{N} = \frac{472}{50} = 9.44 \]

Example 2. Compute the semi-inter-quartile range of the marks of 63 students in Mathematics given below:

<table>
<thead>
<tr>
<th>Marks Group</th>
<th>No. of Students</th>
<th>Marks Group</th>
<th>No. of students</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-10</td>
<td>5</td>
<td>50-60</td>
<td>7</td>
</tr>
<tr>
<td>10-20</td>
<td>7</td>
<td>60-70</td>
<td>3</td>
</tr>
<tr>
<td>20-30</td>
<td>10</td>
<td>70-80</td>
<td>2</td>
</tr>
<tr>
<td>30-40</td>
<td>16</td>
<td>80-90</td>
<td>2</td>
</tr>
<tr>
<td>40-50</td>
<td>11</td>
<td>90-100</td>
<td>0</td>
</tr>
</tbody>
</table>

Solution.

<table>
<thead>
<tr>
<th>Marks Group</th>
<th>Frequency f</th>
<th>Cumulative Frequency c.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-10</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>10-20</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>20-30</td>
<td>10</td>
<td>22</td>
</tr>
<tr>
<td>30-40</td>
<td>16</td>
<td>38</td>
</tr>
<tr>
<td>40-50</td>
<td>11</td>
<td>49</td>
</tr>
<tr>
<td>50-60</td>
<td>7</td>
<td>56</td>
</tr>
<tr>
<td>60-70</td>
<td>3</td>
<td>59</td>
</tr>
<tr>
<td>70-80</td>
<td>2</td>
<td>61</td>
</tr>
<tr>
<td>80-90</td>
<td>2</td>
<td>63</td>
</tr>
<tr>
<td>90-100</td>
<td>0</td>
<td>63</td>
</tr>
</tbody>
</table>

\[ \sum f = 63 \]

To calculate lower Quartile Q1. Here N = 60. So \( \frac{1}{4} (N+1) \)th i.e., 16th students lies in the marks group 20-30. Thus lower quartile class is 20-30.

\[ Q1 = l + \frac{\frac{N-F}{f} \times i}{f} = 20 + \frac{15.75-12}{10} \times 10 = 23.75. \]
Similarly $Q_3 = 40 + \frac{47.25 - 38}{11} \times 10 = 48.4$.

Semi-inter quartile range $= \frac{1}{2} (Q_3 - Q_1)$

$= \frac{1}{2} (48.4 - 23.75) = 12.32$ Marks

### 1.4.4 STANDARD DEVIATION

Root – mean square Deviation :

**Definition.** It is defined as the positive square root of the mean of the squares of the deviations from an origin $A$ and denoted by $s$

$$s = +\sqrt{\left(\frac{1}{N} \sum f(x - A)^2\right)}$$

Mean square deviation. It is defined as the mean of the squares of the deviations from an origin $A$. Thus

$$s^2 = \frac{\sum f(x - A)^2}{n}$$

**Remark.** The origin $A$ may be taken at any arbitrary point $A$.

Standard Deviation :

**Definition.** Standard deviation (or S.D.) is the positive square root of the arithmetic mean of the square deviations of various values from their arithmetic mean $M$. It is usually denoted by $\sigma$. Thus

$$\sigma = s = +\sqrt{\left(\frac{1}{N} \sum f(x - M)^2\right)}$$

**Remarks:**

1. When the deviations is calculated from the arithmetic mean $M$, then root mean square deviation becomes standard deviation.
2. The square of the standard deviation $\sigma^2$ is called variance.
3. The quantity $s^2$ is said to be second moment about the value $A$ and is denoted by $\mu_2$.
4. The variance $\sigma^2$ is called the second moment about the mean $M$ and is denoted by $\mu_2$.

RELATION BETWEEN STANDARD DEVIATION AND ROOT-MEAN SQUARE DEVIATION

Consider a frequency distribution (discrete series)

\[ x : \quad x_1, x_2, \ldots, x_n \]
\[ f : \quad f_1, f_2, \ldots, f_n \]

Let $A$ be the assumed mean and $M$ the arithmetic mean. Also suppose $M-A = d$. Then

\[ \sigma^2 = \frac{1}{N} \sum f (x - M)^2 = \frac{1}{N} \sum f (x - A - M)^2 \]

\[ = \frac{1}{N} \sum f (x - d)^2 \text{ where } X = x - A, d = M - A \]

\[ = \frac{1}{N} \sum f x^2 - 2d \frac{1}{N} \sum f X + d^2 \frac{1}{N} \sum f \]

\[ = \frac{1}{N} \sum f (x - A)^2 - 2d \frac{1}{N} \sum f (x - A) + d^2 \]

\[ = \frac{1}{N} \sum f (x - A)^2 - 2d \left( \frac{1}{N} \sum f x - \frac{1}{N} \sum f A \right) + d^2 \]

\[ = \frac{1}{N} \sum f (x - A)^2 - 2d (M - A) + d^2 \]

\[ = \frac{1}{N} \sum f (x - A)^2 - d^2 = s^2 - d^2 \]

Hence $s^2 = \sigma^2 + d^2 \quad \ldots (1)$

Relation (1) shows that $s$ is least when $d = 0$ i.e., $A = M$ and the least value of $s$ is equal to $\sigma$.

In other words the standard deviation is the least possible root mean square deviation.

Remark. Since $d^2 > 0$, always, therefore, from (1), we have

\[ S^2 > \sigma^2 \]

i.e., mean square deviation about any point $A$ is greater than variance.
SHORT CUT METHOD FOR CALCULATING STANDARD DEVIATION

We know that

$$\sigma = + \sqrt{\left( \frac{1}{N} \sum f\xi^2 - \left( \frac{\sum f\xi}{N} \right) \right)}$$

Where $\xi = (x-A)$ and $A$ = assumed mean

STEP DEVIATION METHOD TO COMPUTE S.D.

If $u = (x-A)/h = \xi/h$, then $\xi = uh$.

By short cut method

$$\sigma^2 = h^2 \left\{ \frac{1}{N} \sum fu^2 - \left( \frac{\sum fu}{N} \right)^2 \right\}$$

$$\sigma = h \sqrt{\left[ \frac{1}{N} \sum fu^2 - \left( \frac{\sum fu}{N} \right)^2 \right]}$$

1.5 ABSOLUTE AND RELATIVE MEASURES OF DISPERSION

The measure of dispersions namely ranger, quartile deviations, inter-quartile deviation. Mean deviation, standard deviation, root mean square deviation (these have been discussed above) are said to be absolute measure of dispersion, since they are expressed in terms of units of observations (Cm., Km., Rs., degree etc.). We know that different units can not be compared; for example a centimeter can not be compared with a rupee. Therefore, the dispersions in different units can not be compared. Also the measures of dispersion depend on the measures of central tendency. Therefore, it is needed to define some measures which are independent of the units of measurement and can be adjusted for measures of central tendency. Such type of measures are called relation measures of dispersion or coefficients of dispersion. These relative measures are pure numbers and are usually expressed as percentages.

They are useful to compare two series in different units and also to compare variations of two series having different magnitudes.

Some of the relative measures of dispersion (or coefficient of dispersion) which are in common use are given below :

(i) Quartile coefficient of dispersion. It is usually denoted by Q.D. and is defined by
Q.D. = \frac{Q_3 - Q_1}{Q_3 + Q_1}

(ii) Coefficient of mean dispersion = Mean deviation about any point 'a' / a.

Here any point 'a' can be replaced by mean, median, mode etc.

(iii) Coefficient of variation or coefficient of dispersion. It is defined by the ration \( \sigma / M \), where \( \sigma \) is standard deviation and \( M \) is the arithmetic mean. It is denoted by C.V. or V. thus

\[
C.V. \text{ or } V = \frac{\sigma}{M}
\]

Sometimes, we define

\[
C.V. \text{ or } V = \frac{\sigma}{M} \times 100
\]

Example 1. Calculate the S.D. and coefficient of variation (C.V.) for the following table:

<table>
<thead>
<tr>
<th>Class</th>
<th>0-10</th>
<th>10-20</th>
<th>20-30</th>
<th>30-40</th>
<th>40-50</th>
<th>50-60</th>
<th>60-70</th>
<th>70-80</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>5</td>
<td>10</td>
<td>20</td>
<td>40</td>
<td>30</td>
<td>20</td>
<td>10</td>
<td>5</td>
</tr>
</tbody>
</table>

Solution. We prepare the following table for the computation of S.D.

<table>
<thead>
<tr>
<th>Class</th>
<th>Mid-value x</th>
<th>f</th>
<th>( u = \frac{x - 35}{10} )</th>
<th>( fu )</th>
<th>( fu^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-10</td>
<td>5</td>
<td>5</td>
<td>-3</td>
<td>-15</td>
<td>45</td>
</tr>
<tr>
<td>10-20</td>
<td>15</td>
<td>10</td>
<td>-2</td>
<td>-20</td>
<td>40</td>
</tr>
<tr>
<td>20-30</td>
<td>25</td>
<td>20</td>
<td>-1</td>
<td>-20</td>
<td>20</td>
</tr>
<tr>
<td>30-40</td>
<td>35</td>
<td>40</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>40-50</td>
<td>45</td>
<td>30</td>
<td>1</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>50-60</td>
<td>55</td>
<td>20</td>
<td>2</td>
<td>40</td>
<td>80</td>
</tr>
<tr>
<td>60-70</td>
<td>65</td>
<td>10</td>
<td>3</td>
<td>30</td>
<td>90</td>
</tr>
<tr>
<td>70-80</td>
<td>75</td>
<td>5</td>
<td>4</td>
<td>20</td>
<td>80</td>
</tr>
</tbody>
</table>

\[ N = \sum f = 140 \]

\[ \sum fu = 65 \]

\[ \sum fu^2 = 385 \]

Let assumed mean = 35 = A (say) and \( h = 10 \)

A.M., \( M = A + h \left( \frac{\sum fu}{N} \right) = 35 + 10 \left( \frac{65}{140} \right) \)

\[ = 35 + 4.64 = 39.64 \]

\[ S.D., \sigma = h \sqrt{\left( \frac{\sum fu^2}{N} - \left( \frac{\sum fu}{N} \right)^2 \right)} \]
\[
= 10 \sqrt{\frac{385}{140} - (.464)^2}
\]

\[
= 10 \sqrt{2.75 - .215} = 10 \sqrt{2.535} = 10 \times 1.59 = 15.9
\]

C.V. = \(\sigma/M \times 100 = 15.9/39.64 \times 100 = 40.11\%\).

### 1.6 MOMENTS

For any frequency distribution, the \(r\)th moment about any point \(A\) is defined as the arithmetic mean of \(r\)th powers of deviations from the point \(A\).

1.7.1 Moments about mean (or Central Moments):

(i) For an individual series. Let \(x_1, x_2, \ldots, x_n\) be the \(n\) values of the variable \(x\), then the \(r\)th moment about the mean (arithmetic mean) \(x\) is denoted by \(\mu_r\) and is defined by

\[
\mu_r = \frac{\sum_{i=1}^{n} (x_i - x)^r}{n}, \text{ for } r = 0, 1, 2, 3, \ldots
\]

(ii) For a frequency distribution. Let

\[
x: \quad x_1 \quad x_2 \quad \ldots \quad x_n
\]

\[
f: \quad f_1 \quad f_2 \quad \ldots \quad f_n
\]

be a discrete frequency distribution. Then the \(r\)th moment \(\mu_r\) about the mean \(x\) is defined by

\[
\mu_r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^r}{N}, \text{ for } r = 0, 1, 2, 3, \ldots \text{ Where } \sum f_i = N
\]

or

Particular Cases.

For \(r = 0\),

\[
\mu_0 = \frac{\sum_{i=1}^{n} f_i(x_i - \bar{x})^0}{N} = \frac{\sum_{i=1}^{n} f_i}{N} = \frac{N}{N} = 1
\]

Hence for all distributions,
\[ \mu_0 = 1 \]

For \( r = 1, \)  \[ \mu_1 = \frac{\sum f_i (x_i - \bar{x})^1}{N} = \frac{1}{N} \sum f_i x_i - \frac{x}{N} \sum f_i = x - \frac{x}{N} = 0. \]

Hence for all distributions,  
\[ \mu_1 = 0 \]

For \( r=2, \)  \[ \mu_2 = \frac{1}{N} \sum f_i (x_i - \bar{x})^2 = \sigma^2 = \text{variance}. \]

Hence for all distributions, we have  
\[ \mu_2 = (\text{standard deviation})^2 = \text{Variance} \]

For \( r=3, \)  \[ \mu_3 = \frac{\sum f_i (x_i - \bar{x})^3}{N}, \text{and so on.} \]

**MOMENTS ABOUT ANY POINT (RAW MOMENTS):**

For any frequency distribution the rth moment about any point \( x = A, \) is defined as the arithmetic mean of the rth powers of the deviations from the point \( x=A \) and is denoted by \( \mu_r \).

If  
\[
\begin{align*}
&X: \quad x_1 \quad x_2 \quad \ldots \quad x_n \\
&F: \quad f_1 \quad f_2 \quad \ldots \quad f_n
\end{align*}
\]
Be discrete frequency distribution, then  
\[ \mu_r = \frac{1}{N} \sum_{i=1}^{n} f_i (x_i - A)^r, \quad r = 0, 1, 2, 3, \ldots, \text{and} \sum_{i=1}^{n} f_i = N. \]

In case of individual series  
\[ \mu_r = \frac{\sum_{i=1}^{n} (x_i - A)^r}{n}, \quad r = 0, 1, 2, 3, \ldots. \]
Particular cases.

For r = 0,

\[ \mu_0' = \frac{1}{N} \sum_{i=1}^{n} f_i(x_i - A)^0 = \frac{1}{N} \sum_{i=1}^{n} f_i = \frac{N}{N} = 1 \]

For r=1

\[ \mu_1' = \frac{1}{N} \sum_{i=0}^{n} f_i(x_i - A)^1 = \frac{1}{N} \sum_{i=1}^{n} f_i x_i - \frac{1}{N} A \sum_{i=1}^{n} f_i \]

\[ = x - A/N \times N = x - A = d \text{ (say)} \]

Where x = Arithmetic mean of given data.

For r=2

\[ \mu_2' = \frac{1}{N} \sum_{i=1}^{n} f_i(x_i - A)^2 = \frac{1}{N} \sum_{i=1}^{n} f_i x_i - \frac{1}{N} A \sum_{i=1}^{n} f_i \]

\[ = \frac{1}{N} \sum_{i=1}^{n} f_i (x_i - \bar{x})^2 + \frac{(x - A)^2}{N} \sum_{i=1}^{n} f_i \]

\[ = \sigma^2 + (x-A)^2 = \sigma^2 + d^2. \]

For r=3

\[ \mu_3' = \frac{1}{N} \sum_{i=1}^{n} f_i(x_i - \bar{x})^3, \text{ and so on.} \]

RELATION BETWEEN CENTRAL MOMENTS (\(\mu_r\)) AND MOMENTS ABOUT ANY POINT (\(\mu_r'\)):

We have:

\[ \mu_r = \mu_r' - C_1 \mu_1' \mu_{r-1}' + C_2 (\mu_1')^2 \mu_{r-2}' - \ldots + (-1)^r (\mu_r') \]

MOMENTS ABOUT THE ORIGIN:

If

\[ x : \quad x_1 \quad x_2 \quad \ldots \quad x_n \]
\[ f: \quad f_1 \quad f_2 \quad \ldots \quad f_n \]
be a discrete frequency distribution, then the rth moment about the origin is denoted by $V_r$, (say) and is defined by

$$V_r = \frac{1}{N} \sum_{i=1}^{n} f_i x_i^r, \text{ r}=0,1,2,3,... \text{ and } \sum_{i=1}^{n} f_i = N$$

Putting $r=0,1,2,...$, we get

$$V_0 = \frac{1}{N} \sum_{i=1}^{n} f_i = \frac{N}{N} = 1$$

$$V_1 = \frac{1}{N} \sum_{i=1}^{n} f_i x_i^r$$

$$= \frac{1}{N} \sum_{i=1}^{n} f_i \{(x_i - \bar{x}) + \bar{x}\}^r$$

$$= \frac{1}{N} \sum_{i=1}^{n} f_i \{(x_i - \bar{x})^r + r C_1 (x_i - x)^{r-1} x + r C_2 (x_i - x)^{r-2} (x)^2 + ... + (x)^r\}$$

$$V_r = \mu^r + r C_1 \mu_{r-1} \bar{x} + r C_2 \mu_{r-2} (\bar{x})^2 + ... + (\bar{x})^r$$

### 1.7 KARL PEARSON’S $\beta$ AND $\gamma$ COEFFICIENTS

Karl Pearson gave the following four coefficients. Calculated from the central moments, which are defined as

<table>
<thead>
<tr>
<th>Beta coefficients</th>
<th>Gamma coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1 = \mu_3^2/\mu_2^3$</td>
<td>$\gamma_1 = \pm \sqrt{\beta_1}$</td>
</tr>
<tr>
<td>$\beta_2 = \mu_4/\mu_2^2$</td>
<td>$\gamma_2 = \beta_2 - 3 = \frac{\mu_4 - 3 \mu_2^2}{\mu_2^2}$</td>
</tr>
</tbody>
</table>

The sign of $\gamma_1$ depends upon $\mu_3$ is positive then $\gamma_1$ is positive. If $\mu_3$ is negative then $\gamma_1$ is negative. The above four coefficients are pure numbers and thus do not have any unit. The $\beta$ and $\gamma$ coefficients give some idea about the shape of the curve obtained from the frequency distribution. This we shall discuss in the topic Kurtosis and Skewness.

Example 1. Calculate the first four central moments from the following data

<table>
<thead>
<tr>
<th>Class</th>
<th>0-10</th>
<th>10-20</th>
<th>20-30</th>
<th>30-40</th>
<th>40-50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>4</td>
</tr>
</tbody>
</table>
Solution. Let $A=25$. Here $h=10$. To facilitate the calculations, let

$$u = \frac{x - A}{h} = \frac{x - 25}{10}$$

<table>
<thead>
<tr>
<th>Class</th>
<th>Frequency $f$</th>
<th>Mid-value $x$</th>
<th>$U = \frac{x - 25}{10}$</th>
<th>$fu$</th>
<th>$fu^2$</th>
<th>$fu^3$</th>
<th>$fu^4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-10</td>
<td>1</td>
<td>5</td>
<td>-2</td>
<td>-2</td>
<td>4</td>
<td>-8</td>
<td>16</td>
</tr>
<tr>
<td>10-20</td>
<td>3</td>
<td>15</td>
<td>-1</td>
<td>-3</td>
<td>3</td>
<td>-3</td>
<td>3</td>
</tr>
<tr>
<td>20-30</td>
<td>5</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>30-40</td>
<td>7</td>
<td>35</td>
<td>1</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>40-50</td>
<td>4</td>
<td>45</td>
<td>2</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>$N = \sum f = 20$</td>
<td></td>
<td>$\sum fu = 10$</td>
<td>$\sum fu^2 = 30$</td>
<td>$\sum fu^3 = 28$</td>
<td>$\sum fu^4 = 90$</td>
<td></td>
</tr>
</tbody>
</table>

The $r$th moment about any point $x=A=25$ is given by

$$\mu_r = \frac{h^r \sum fu^r}{N}$$

$$\mu_1 = \frac{h^1 \sum fu^1}{N} = 10 \times 10 / 20 = 5$$

$$\mu_2 = \frac{h^2 \sum fu^2}{N} = 100 \times 30 / 20 = 150$$

$$\mu_3 = \frac{h^3 \sum fu^3}{N} = 1000 \times 28 / 20 = 1400$$

$$\mu_4 = \frac{h^4 \sum fu^4}{N} = 10000 \times 90 / 20 = 45000.$$  

To calculate central moments,

$$\mu_1 = 0 \text{ (always)}$$

$$\mu_2 = \mu_2^* - (\mu_1^*)^2 = 150 - (5)^2 = 125$$

$$\mu_3 = \mu_3^* - 3 \mu_2^* \mu_1^* + 2 (\mu_1^*)^3 = 1400 - 3 \times 150 \times 5 + 2 (5)^3 = -600$$

$$\mu_4 = \mu_4^* - 4 \mu_3^* \mu_1^* + 6 \mu_2^* (\mu_1^*)^2 - 3 (\mu_1^*)^4$$

$$= 45000 - 4 \times 1400 \times 5 + 6 \times 150 \times (5)^2 - 3 (5)^4 = 37625$$
1.8. SKEWNESS

By skewness in some frequency distribution, we mean the lack in symmetry. [If the frequencies are symmetrically distributed about the mean, then the distribution is called symmetrical, in other words, a distribution is called symmetrical when the values equidistant from the mean have equal frequencies.] Skewness is also termed as asymmetry.

Skewness denotes the tendency of distribution to depart from symmetry. According to Simpson, “Skewness or asymmetry is the attribute of a frequency distribution that extends further on one side of the class with the highest frequency that on the other”.

We know that for a symmetrical distribution the mean, median and mode coincide. Therefore, skewness in a distribution is shown when these three averages do not coincide. Skewness indicates that the frequency curve has a longer tail on one side for the average. When the frequency curve has a longer tail on right side, the skewness is called positive. When the frequency curve has a longer tail on left side, the skewness is called negative. In other words, the skewness is positive if \( M_0 < M_d < M \) and negative if \( M < M_d < M_0 \), where \( M \), \( M_d \) and \( M_0 \) are mean, median and mode respectively.

MEASURE OF SKEWNESS:

We shall give following three measures to measure the skewness:

(i) First coefficient of skewness. It is also known as Bowley’s coefficient of skewness and is defined as

\[
\text{Coefficient of skewness} = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1} = \frac{(Q_1 - M_d) - (M_d - Q_1)}{(Q_3 - M_d) + (M_d - Q_1)} = J_Q.
\]

Where \( Q_1 \) and \( Q_3 \) are lower and upper quartiles respectively and \( M_d \) is median. Clearly this measure is based on the fact that in a skew curve, the median does not lie half way between \( Q_1 \) and \( Q_3 \). This formula for coefficient of skewness is used when mode is well defined.
Second coefficient of skewness. It is also called Karl Pearson’s coefficient of skewness and is defined as

\[
\text{Coefficient of skewness} = \frac{\text{Mean} - \text{Mode}}{S \tan\text{dardDeviation}} = \frac{M - M_0}{\sigma} = J
\]

If mode is not well defined, then

\[
\text{Coeff. Of skewness} = \frac{3(M - M_0)}{\sigma}
\]

Clearly this measure is based on the fact that mean and mode are not coincident.

Note that both of the above coefficients are pure numbers since both the numerator and denominator have the same dimensions.

(ii) Coefficients of Skewness Based on Moments. When there is symmetrical distribution, all the moments of odd order about the arithmetic mean (i.e., \(\mu_1, \mu_3, \mu_5\) etc.) vanish. If the values of these coefficients do not vanish then there is skewness in the frequency distribution.

According to Karl Pearson the coefficients of skewness are exactly given by the following formulae:

First Coefficient of Skewness = \(\frac{\mu_3}{\sqrt{\mu_2^3}} = \sqrt{\beta_1} = \gamma_1\)

Second Coefficient of Skewness = \(\frac{\sqrt{\beta_1(\beta_2 + 3)}}{2(5\beta_2 - 6\beta_1 - 9)}\)

If skewness in the series is very small then second coefficient of skewness should be used.

Coefficients of skewness based on moments are also called Moment Coefficient of skewness.

Examples 1. Compute the Bowley’s coefficient of skewness for the following frequency distribution:

<table>
<thead>
<tr>
<th>Marks</th>
<th>0-10</th>
<th>10-20</th>
<th>20-30</th>
<th>30-40</th>
<th>40-50</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of students</td>
<td>2</td>
<td>7</td>
<td>10</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>

Solution.

<table>
<thead>
<tr>
<th>Marks</th>
<th>Frequency</th>
<th>C.F.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-10</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>10-20</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>20-30</td>
<td>10</td>
<td>19</td>
</tr>
</tbody>
</table>
Lower quartile, \( Q_1 = \frac{N}{4} \) th term = \( \frac{27}{4} \) th term = 6.75 th term.  
\[
Q_1 = l + \frac{\frac{1}{4} N - F}{f} \times i = 10 + \frac{6.75 - 2}{7} \times 10 
\]
\[
Q_1 = 10 + \frac{4.75}{7} \times 10 = 10 + 6.79 = 16.79 \text{ marks.}
\]

Median, \( \text{Md} = \frac{N}{2} \) th term = 13.5 th term  
\[
\text{Md} = l + \frac{\frac{1}{4} N - F}{f} \times i = 20 + \frac{13.5 - 9}{10} \times 10 = 20 + 4.5 = 24.5 \text{ marks.}
\]

\( Q_3 = 3 \times \left( \frac{N}{4} \right) \) th term = 20.25 th term.  
\[
Q_3 = l + \frac{\frac{3}{4} N - F}{f} \times i 
\]
\[
Q_3 = 30 + \frac{20.25 - 19}{5} \times 10 = 30 + 2.5 = 32.5 \text{ marks.}
\]

Bowley’s coefficient of skewness  
\[
JQ = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1} 
\]
\[
= \frac{32.5 + 16.79 - 2 \times 24.5}{32.5 - 16.79} = \frac{0.29}{15.79} = 0.02.
\]

1.9 KURTOSIS

In Greek language kurtosis means ‘bulgines.’ kurtosis indicates the nature of the vertex of the curve. Several statisticians defined kurtosis. Some of these definitions are:

“In statistics, kurtosis refers to the degree of flatness of peaked ness in the region about the mode of frequency curve. The degree of kurtosis of a distribution is measured relative to the peaked ness of normal curve.”

“A measure of kurtosis indicates the degree to which a curve of the frequency distribution is peaked or flat-topped.”

Karl Pearson in 1905 defined following three types of curves:

1. Normal Curve or Measokurtic Curve. A curve which is neither flat nor peaked is called a normal curve or meso-kurtic curve. For such type of curve we have \( \beta_2 = 3 \) and \( \gamma^2 = 0 \).
2. Leptokurtic Curve. A curve which is more peaked than the normal curve is called leptokurtic curve. For such type of curve, we have $\beta_2>3$ and $\gamma_2>0$.

3. Platykurtic Curve. A curve which is more flatter than the normal curve is called platykurtic curve. For such type of curve, we have $\beta_2<3$ and $\gamma_2<0$.

Measure of Kurtosis

Second and fourth moments are used to measure kurtosis. Karl Pearson gave the following formula to measure kurtosis:

$$Kurtosis = \beta_2 = \frac{\mu_4}{\mu_2^2}.$$

To measure kurtosis, $\gamma_2$ is used and it is given by the following formula:

$$\gamma_2 = \beta_2 - 3 = \frac{\mu_4 - 3\mu_2^2}{\mu_2^2}.$$

Deductions.

(1) If $\gamma_2 = 0$, the curve is normal.
(2) If $\gamma_2>0$, the curve is leptokurtic.
(3) If $\gamma_2<0$, the curve is platykurtic.

Example 2. The fourth moment about mean of frequency distribution is 768. What must be value of its standard deviation in order that the distribution be

(i) Leptokurtic
(ii) Mesokurtic
(iii) Platykurtic.

Given $\mu_4 = 768$.

Kurtosis = $\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4} = \frac{768}{\sigma^4}$

Now the distribution will be

1. Leptokurtic if $\beta_2 > 3 \Rightarrow \frac{768}{\sigma^4} > 3 \Rightarrow \sigma^4 < \frac{768}{3}$
   
   $\Rightarrow \sigma^4 < 256 \Rightarrow \sigma^4 < (4)^4 \Rightarrow \sigma < 4$

2. Mesokurtic if $\beta_2 = 3 \Rightarrow \frac{768}{\sigma^4} = 3 \Rightarrow \sigma = 4
3. **Platykurtic if** $\beta_2 < 3 \Rightarrow \sigma^4 < 3 \Rightarrow \sigma > 4$

---

### 1.10 Let us sum up

After going through this unit, you would achieved the objectives stated earlier in the unit. Let us recall what we have discussed so far –

* An average is a single value within the range of the data that is used to represent all the values in the series.

* To find the arithmetic mean, add the values of all terms and then divide sum by the number of terms, the quotient is the arithmetic mean.

* The **median** is the value of the variable which divides the group into two equal parts one part comprising all values greater, and the other all values less than the median.

* The **mode** is that value (or size) of the variate for which the frequency is maximum or the point of maximum frequency or the point of maximum density. In other words, the mode is the maximum ordinate of the ideal curve which gives the closest fit to the actual distribution.

* For moderately asymmetrical distribution (or for asymmetrical curve), the relation

  \[
  \text{Mean} - \text{Mode} = 3 (\text{Mean} - \text{Median}),
  \]

  approximately holds. In such a case, first evaluate mean and median and then mode is determined by

  \[
  \text{Mode} = 3 \text{ Median} - 2 \text{ Mean}.
  \]

  If in the asymmetrical curve the area on the left of mode is greater than area on the right then

  \[
  \text{Mean} < \text{median} < \text{mode}, \text{ i.e., } (M < Md < M0)
  \]

* If $f_1, f_2, \ldots, f_n$ are the frequencies of $x_1, x_2, \ldots, x_n$ respectively, then geometric mean $G$ is given by

  \[
  G = \left\{ x_1^{f_1} x_2^{f_2} \ldots x_n^{f_n} \right\}^{1/N}
  \]

  \[
  N = f_1 + f_2 + \ldots + f_n
  \]
* If \( f_1, f_2, \ldots, f_n \) be the frequencies of \( x_1, x_2, \ldots, x_n \) (none of them being zero) then harmonic mean \( H \) is given by
\[
H.M. = \frac{\sum f}{\sum f \times \frac{1}{x}}
\]

* The values of the variate which divide the total frequency into four equal parts, are called quartiles

* The values of the variate which divide the total frequency into ten equal parts are called deciles

* The values of the variate which divide the total frequency into hundred equal parts, are called percentiles.

* The range of a set of numbers (data) is the difference between the largest and the least numbers in the set.

* Semi-inter quartile range (or quartile deviation) is denoted by \( Q \) and is defined by
\[
Q = (Q_3 - Q_1)/2
\]

* If \( x_1, x_2, \ldots, x_k \) occur with frequencies \( f_1, f_2, \ldots, f_k \) respectively, then the mean deviation (\( \delta_m \)) is defined by
\[
\delta_m = \frac{1}{N} \sum_{j=1}^{k} f_j (x_j - M) = \frac{1}{N} \sum f (x - M)
\]

* Standard deviation (or S.D.) is the positive square root of the arithmetic mean of the square deviations of various values from their arithmetic mean \( M \). It is usually denoted by \( \sigma \). Thus
\[
\sigma = s = \sqrt{\left( \frac{1}{N} \sum f (x - M)^2 \right)}
\]

* For any frequency distribution, the \( r \)th moment about any point \( A \) is defined as the arithmetic mean of \( r \)th powers of deviations from the point \( A \).

* Skewness denotes the tendency of distribution to depart from symmetry.
* measure of kurtosis indicates the degree to which a curve of the frequency distribution is peaked or flat-topped.

1.11 Check your progress : The key

Exercise

Q.1) Calculate the measure of Kurtosis for the following distribution.

<table>
<thead>
<tr>
<th>Marks</th>
<th>3-15</th>
<th>15-25</th>
<th>25-35</th>
<th>35-45</th>
<th>45-55</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of candidate</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>4</td>
</tr>
</tbody>
</table>

Q.2) The first four moments of the distribution about the value 5 of a variable are 2, 20, 40 and 50. Find the central moment.

Unit-2

Probability

Structure:

2.0 Introduction

2.1 Objectives

2.2 Random experiment

2.3 Sample space
2.4 Event

2.5 Additional law of probability

2.6 Conditional probability

2.7 Multiplicative law of probability

2.8 Independent and dependent event

2.9 Bay’s theorem

2.10 Random variable

2.11 Probability density function

2.12 Continuous probability distribution

2.13 Cumulative distribution function

2.14 Mean, Median, Mode and Moments for a continuous distribution function

2.15 Mathematical expectation

2.16 Covariance

2.17 Let us sum up

2.18 Check your progress : The key
2.0 INTRODUCTION

If a coin is tossed repeatedly under essentially homogeneous and similar conditions, then one is not sure if a ‘head’ or a ‘tail’ will be obtained. Such types of phenomena (i.e. phenomena which do not lend itself to deterministic approach) are called unpredictable or ‘probabilistic’ phenomena.

In 1993 A. N. Komogrov, a Russian mathematician, tried successfully to relate the theory of probability with the set theory by axiomatic approach. The axiomatic definition of probability includes both the classical and the statistical definitions as particular cases and overcomes the definitions of each of them.

2.1 OBJECTIVES

The main aim of this unit is to study the probability. After going through this unit you should be able to:

- describe random experiments, sample space, additive law and multiplicative law of probability, dependent and independent events etc;
- calculate mean, mode, median, and moments for a continuous distribution function, mathematical expectation;
- find out covariance;
- know Bay’s theorem, random variables, probability density function, continuous probability distribution, cumulative distributive function, etc.
2.2 RANDOM EXPERIMENT

Consider a bag containing 4 white and 5 black balls. Suppose 2 balls are drawn at random. Here the natural phenomenon is that ‘both balls may be white’ or ‘one white and one black’ or ‘both black’. Thus there is a probabilistic situation.

We feel intuitively in the following statements.

(i) The probability of getting a ‘tail’ in one toss of an unbiased coin is $\frac{1}{2}$.
(ii) The probability of getting an ‘ace’ in a single of an unbiased die is $\frac{1}{6}$.

Similarly the probability of getting a 2 or 3 in a single throw of an unbiased die should be the sum of probabilities of getting a 2 or 3 i.e. $\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$. In other words, there should be the sum of probabilistic situations, we need mathematical models.

Any probabilistic situation is called a random experiment and is denoted by $E$. Each performance in random experiment is called a trial and the result of a trial is called an ‘outcome’ or a ‘sample point’ or an ‘elementary event’.

2.3 SAMPLE SPACE.
A sample space of a random experiment is the set of all possible outcomes of that experiment and is denoted by S.

Finite sample space. A sample space containing finite number of sample points, is called a finite sample space.

2.4 EVENT.

Of all the possible outcomes in the sample space of an experiment some outcomes satisfy a specified description, it is called an event. In other words ‘every non-empty subset of a sample space is called an event of the sample space’. It is denoted by E. Several events are denoted by \( E_1, E_2, \) etc.

Certain and impossible events. If S is a sample space, then S and \( \Phi \) are both subsets of S and so S and \( \Phi \) both are events. S is called certain event and \( \Phi \) is called impossible event.

Equally likely events. Two events are considered equally likely if one of them cannot be expected in preference to the other.

For example, if an unbiased coin is tossed then we may get any of head (H) or tail (T), thus the two different events are equally likely.
Exhaustive Events. All possible outcomes in a trial, are called exhaustive events.

For example, if an unbiased die is rolled, then we may obtain any one of the six numbers 1, 2, 3, 4, 5 and 6. Hence there are six exhaustive events in this trial.

Favorable events. The total number of favorable outcomes (or ways) in a trial, to happen an event, are called favorable events.

For example, If a pair of fair dice is tossed then the favorable events to get the sum 7 are six:

\[(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\].

Mutually Exhaustive or Incompatible Events. Two or more than two events are called mutually exclusive events if there is no element (or outcome or result) common to these events. In other words, events are called mutually exclusive if the happening of one of them prevents or precludes the happening of the other events.

If \(E_1\), \(E_2\) are two mutually exclusive events then \(E_1 \cap E_2 = \emptyset \Rightarrow E_1\) and \(E_2\) are mutually exclusive.

EXAMPLES ON SAMPLE SPACE AND EVENT
Example 1. In a single toss of a fair die, find (a) sample space (b) event of getting an even number (c) event of getting an odd number (d) event of getting numbers greater than 3, (e) event of getting numbers less than 4.

Solution. (a). When we toss a die, then we may get any of the six numbers, 1, 2, 3, 4, 5, and 6. hence the set of these six numbers is the sample space $S$ for this experiment, i.e.,

$$S = \{1, 2, 3, 4, 5, 6\};$$

In the above experiment, to get an even number is an event, say $E_1$; to get an odd number is an event, say $E_2$; to get numbers greater than 3 is an event, say $E_3$; and to get numbers less than 4 is an event, say $E_4$. Thus

(b) $E_1 = \{2, 4, 6\}$

(c) $E_2 = \{1, 3, 5\}$

(d) $E_3 = \{4, 5, 6\}$

(e) $E_4 = \{1, 2, 3\}$

Example 2. Consider an experiment in which two coins are tossed together. Find the sample space. Find also the following events:

(a) Heads on the upper faces of coins, (b) head on one and tail on other, (c) Tails on both, (d) at least one head.

Solution. If $H$ denotes ‘head’ and $T$ denotes ‘tail’ then the toss of two coins can lead to four cases $(H, H), (T, T), (H, T), (T, H)$ all equally likely. Hence the sample space $S$ is the set of all these four ordered pairs, thus

$$S = \{(H, H), (T, T), (H, T), (T, H)\};$$
In this experiment let $E_1, E_2, E_3$ and $E_4$ be the events of getting both heads, one head and one tail, both tails and at least one head respectively, then

(a) $E_1 = \{(H,H)\}$

(b) $E_2 = \{(H,T),(T,H)\}$

(c) $E_3 = \{(T,T)\}$

(d) $E_4 = \{(H,H),(H,T),(T,H)\}$

Simple And Compound Events

Consider a random experiment and let $e_1, e_2, \ldots, e_n$ be the outcomes or sample points so that the sample space $S$ for this experiment is given by $X = \{e_1, e_2, \ldots, e_n\}$.

Let $E$ be an event related to this experiment then $E \subseteq S$. The set $E$ representing the event, may have only one or more elements of $S$. Based upon this fact, every event can be divided into following two types of events:

Simple event. If $E$ contains only one element of the sample space $S$, then $E$, is called simple event. Thus

$$E = \{e_i\}$$

where $e_i \in S$, is a simple event since it contains only one element of $S$.

Composition of events
Following are three fundamental rules to composite two or more events by the help of set notations.

Let $S$ be a sample space and $A$ and $B$ be its any two events:

(i) The event represented by $A \cup B$ or $A+B$. If the event $E$ happens when $A$ happens or $B$ happens then $E$ is denoted by $A \cup B$ i.e., the event $E$ represented by $A \cup B$ includes all those elements (or outcomes or results) which $A$ or $B$ contain.

\[ A \cup B \]

(ii) The event represented by $A \cap B$ or $AB$. If the event $E$ happens when the events $A$ and $B$ both happen then the event $E$ is represented by $A \cap B$ i.e., the event $E$ denoted by $A \cap B$ includes elements (or outcomes) common to both $A$ and $B$. The shaded area in figure represents the event $E = A \cap B$.

\[ A \cap B \]
(iii) Complement of event A or the event $\overline{A}$ or $\hat{A}$. If the event E happens when the event A does not happen then E is denoted by $\overline{A}$.

2.5 ADDITIVE LAW OF PROBABILITY

Theorem. If $E_1$ and $E_2$ are any two events then

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2).$$

Proof. Let $S$ be the sample space and $n$ be the number of elements in $S$. Let $l$ be the number of elements in $E_1$ and $m$ the number of elements in the event $E_2$, i.e.,

$$n(S) = n, \quad n(E_1) = l, \quad n(E_2) = m.$$

if the events $E_1$ and $E_2$ are not mutually exclusive then $E_1 \cap E_2 \neq \emptyset$.

Let $n(E_1 \cap E_2) = r$
Clearly \( n(E_1 \cup E_2) = l + m - r \).

Now the probability of \( E_1 \) or \( E_2 \) happening

\[
P(E_1 \cup E_2) = \frac{n(E_1 \cup E_2)}{n(S)} = \frac{l + m - r}{n}
\]

\[
= \frac{n(E_1)}{n(S)} + \frac{n(E_2)}{n(S)} - \frac{n(E_1 \cap E_2)}{n(S)}
\]

or

\[
P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)
\]

(1)

cor. If \( E_1 \) and \( E_2 \) be mutually exclusive events then.

\[
E_1 \cap E_2 = \emptyset \quad \text{and} \quad n(E_1 \cap E_2) = 0
\]

Now from (1), we have

\[
P(E_1 \cup E_2) = P(E_1) + P(E_2)
\]

Example 1. If \( \frac{1}{4} \) is the probability of winning a race by the horse A and \( \frac{1}{3} \) be the probability of winning the same race by the horse B. Find the probability that one of these horse will win.
Solution. Let $E_1$ and $E_2$ be the events that the horse A and B wins the race respectively. Then

$$P(E_1) = \frac{1}{4}, \ P(E_2) = \frac{1}{3}$$

We know that if the horse A wins the race then the horse B cannot win the race and if B wins the race then A can not win. Hence the events $E_1$ and $E_2$ are mutually exclusive events. Therefore, the probability that any one of A or B ins the race is given by

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) = \frac{1}{4} + \frac{1}{3} = \frac{7}{12}$$

Example 2. Discuss and criticize the following:

$$P(A) = \frac{1}{6}, \ P(B) = \frac{1}{4}, \ P(C) = \frac{2}{3}$$

for the probabilities of three mutually exclusive events A,B,C.

Solution. Since A,B, C are mutually exclusive events, therefore:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) = \frac{1}{6} + \frac{1}{4} + \frac{2}{3} = \frac{13}{12} > 1$$

which is impossible
2.6 CONDITIONAL PROBABILITY.

When the happening of an event $E_1$ depends upon the happening of an another event $E_2$ then the probability of $E_1$ is called conditional probability and is denoted by $P(E_1/E_2)$. Thus $P(E_1/E_2)$ denotes the conditional probability for the event $E_1$ when the $E_2$ has already happened.

2.7 MULTIPLICATIVE LAW OF PROBABILITY

Theorem. If $E_1$ and $E_2$ are two events, the respective probabilities of which are known, then the probability that both will happen simultaneously is the product of probability of $E_1$ and the conditional probability of $E_2$ when $E_1$ has already occurred i.e.,

$$P(E_1 \cap E_2) = P(E_1)P(E_2/E_1).$$

Proof. Let $S$ be the sample space for an experiment and $E_1$ and $E_2$ be its two events.

Suppose the event $E_1$ has occurred and $E_1 \neq \Phi$. Since $E_1 \subset S$ and the event $E_1$ has occurred, therefore all elements of $S$ cannot occur and only those elements of $S$ which belong to $E_1$ can occur. In this case the reduced sample space will be $E_1$. Now if the event $E_2$ occurs, then all elements of $E_2$ cannot occur but only those elements of $E_2$ which belong to $E_1$ can occur. The set of common elements is $E_1 \cap E_2$. Hence the probability of $E_2$ when $E_1$ has occurred i.e., $P(E_2/E_1)$ [i.e., conditional probability of $E_2$ when $E_1$ has occurred] is given by
\[ P(E_2/E_1) = \frac{n(E_1 \cap E_2)}{n(E_1)} = \frac{n(E_1 \cap E_2)}{n(S)} \]

\[ = \frac{P(E_1 \cap E_2)}{P(E_1)} \]

\[ \therefore P(E_1 \cap E_2) = P(E_1)P(E_2/E_1). \]

Similarly we can prove that

\[ P(E_1 \cap E_2) = P(E_2)P(E_1/E_2). \]

2.8 INDEPENDENT AND DEPENDENT EVENTS

Definition. Let \( E_1 \) and \( E_2 \) be any two events of a sample space. If the occurrence of \( E_1 \) does not depend on the occurrence of \( E_2 \) and the occurrence of \( E_2 \) does not depend on the occurrence of \( E_1 \) or in other words the occurrence of any one does not depend on the occurrence of other then \( E_1 \) and \( E_2 \) are called independent events otherwise they are called dependent events.

Example1. if \( A \) and \( B \) are two events, where

\[ P(A) = \frac{1}{2}, P(B) = \frac{1}{3}, \text{and } P(A \cap B) = \frac{1}{4}, \text{then evaluate the following:} \]

\[ a. ~ P(A/B) \]

\[ b. ~ P(B/A) \]
c. \( P(A \cup B) \)

Solution.

a. \( P(A \cap B) = P(B) \cdot P(A / B) \)

\[
P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{1}{4} = \frac{3}{4}
\]

b. \( P(A \cap B) \)

\[
P(A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)} = \frac{1}{4} = \frac{2}{4} = \frac{1}{2}
\]

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

\[
P(A \cup B) = \frac{1}{2} + \frac{1}{3} - \frac{1}{4} = \frac{7}{12}
\]

2.9 BAYE’S THEOREM

An event \( B \) can be explained by a set of exhaustive and mutually exclusive hypothesis \( A_1, A_2, \ldots, A_n \). Given ‘a priori’ probabilities \( P(A_1), P(A_2), \ldots, P(A_n) \) corresponding to a total absence of knowledge regarding the occurrence of \( B \) and conditional probabilities
\[ P(B/A_1), P(B/A_2), \ldots, P(B/A_n) \]

1. The ‘a posterior’ probability \( P(A_j/B) \) of some event \( A_j \) is given by

\[
P(A_j/B) = \frac{P(A_j)P(B/A_j)}{\sum_{i=1}^{n} P(A_i)P(B/A_i)}
\]

2. The probability \( P(C/B) \) of materialization of another event \( C \) is given by

\[
P(A_j/B) = \frac{\sum_{i=1}^{n} P(A_i)P(B/A_i)P(C/A_iB)}{\sum_{i=1}^{n} P(A_i)P(B/A_i)}
\]

Proof: (a) Since the event \( B \) can occur when either \( A_1 \) occurs, or \( A_2 \) occurs, or \( \ldots, A_n \) occurs i.e., \( B \) can occur in composition with either \( A_1 \) or \( A_2 \) consequently

\[ B = BA_1 \cup BA_2 \cup BA_3 \cup \ldots \ldots BA_n \]

\[
P(B) = P(BA_1 \cup BA_2 \cup BA_3 \cup \ldots \ldots BA_n)
\]

Since \( A_1, A_2, \ldots, A_n \) are mutually exclusive, hence \( BA_1, BA_2, \ldots, BA_n \) are mutually exclusive forms, therefore by total probability theorem, we have
\[ P(B) = P(BA_1) + P(BA_2) + P(BA_3) + \ldots + P(BA_n) \]

\[ = \sum_{i=1}^{n} P(BA_i) = \sum_{i=1}^{n} P(A_i)P(B / A_i) \]

Where \( P(B / A_i) \) is the conditional probability of \( B \) when \( A_i \) has already occurred.

**Now from the theorem of compound probability, we have**

\[ P(A_jB) = P(A_j)P(A_j / B) \]

\[ P(A_j / B) = \frac{P(A_jB)}{P(B)} = \frac{P(A_j)P(A_j / B)}{P(B)} \quad \ldots(2) \]

*From (1) and (2) we get*

\[ P(A_j / B) = \frac{P(A_j)P(B / A_j)}{\sum_{i=1}^{n} P(A_i)P(B / A_i)} \quad \ldots(3) \]

(b) the further event \( C \) can occur in \( n \) mutually exclusive cases namely \( A_1C/B, \ldots, A_nC/B. \) hence the conditional probability of \( C \) is given by

\[ P(C / B) = P(A_1C / B) + P(A_2C / B) + \ldots + P(A_nC / B) \]

\[ = \sum_{i=1}^{n} P(A_iC / B) = \sum_{i=1}^{n} P(A_i / B)P(C / A_iB) \]
\[
\sum_{i=1}^{n} P(A_i)P(B/A_i)P(C/A,B) \\
= \frac{\sum_{i=1}^{n} P(A_i)P(B/A_i)}{\sum_{i=1}^{n} P(A_i)}
\]

Example 1. A bag contains 3 white and 2 black balls, another bag contains 5 white and 3 black balls. If a bag is selected at random and a ball is drawn from it, find the probability that it is white.

Solution. Let B be the event of getting one white ball, and \(A_1, A_2\) be the events of choosing first bag and second bag respectively.

\[P(A_1) = \text{the probability of selecting the first bag out of two bags} = \frac{1}{2}\]

Similarly, \(P(A_2) = \text{the probability of selecting second bag} = \frac{1}{2}\)

\[P(B/A_1) = \text{the conditional probability of drawing one white ball while first bag has been selected} = \frac{3\binom{5}{1}}{\binom{8}{1}} = \frac{3}{5}\]

Similarly \(P(B/A_2) = \frac{5}{8}\)

\[\therefore \quad P(B) = \text{The probability that a white ball is drawn}\]
\[ \sum_{i=1}^{2} P(A_i)P(B \mid A_i) \]
\[ = P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2) \]
\[ = \frac{1}{2} \cdot \frac{3}{5} + \frac{1}{2} \cdot \frac{5}{8} = \frac{49}{80} \]

Ans.

2.10 RANDOM VARIABLES.

Definition. A real valued function defined on a sample-space is called a random-variable (or a discrete random variable).

A random variable can assume only a set of real values and the values which the variable takes depends on the chance. Random variable is also called stochastic variable or simply a variate. For example.

Suppose a perfect die is thrown then x, the number of points on the die is random variable since x has the following two properties:

(i) x takes only a set of discrete values 1, 2, 3, 4, 5, 6;
(ii) the values which x takes depends on the chance.

Actually x takes values 1, 2, 3, 4, 5, 6 each with probability 1/6.

The set of values 1, 2, 3, 4, 5, 6 with their probabilities 1/6 is called the Probability Distribution of variate x.

In general suppose that corresponding to x exhaustive and mutually exclusive cases obtained from a trail, a variate x takes n values \( x_1, x_2, \ldots, x_n \) with their probabilities \( p_1, p_2, \ldots, p_n \).
The set of values $x_i$ (for $i=1,2\ldots n$) with their probabilities $p_i$ (for $i=1,2\ldots n$) is called the Probability Distribution of the variable of that trial. It is to be noted that most of the properties of frequency distribution will be equally applicable to probability distribution.

**Continuous Variate**

So far we have discussed with discrete variate which takes a finite set of values.

When we deal with variates like weights and temperature then we know that these variates can take an infinite number of values in a given interval. Such type of variates are known as continuous variates.

Definition. A variate which is not discrete, i.e., which can take infinite number of values in a given interval $a \leq x \leq b$, is called a continuous variate.

---

2.11 PROBABILITY DENSITY FUNCTIONS

Let $X$ be a continuous random variable and let the probability of $X$ falling in the infinitesimal interval $(x - \frac{1}{2}dx, x + \frac{1}{2}dx)$ be expressed by $f(x)dx$ i.e.

$$P(x - \frac{1}{2}dx < X < x + \frac{1}{2}dx) = f(x)dx,$$

where $f(x)$ is a continuous function of $X$ and satisfies the following two conditions:

(i) $f(x) \geq 0$
(ii) \[ \int_{a}^{b} f(x) \, dx = 1, \text{ if } a \leq X \leq b \]
\[ \int_{-\infty}^{\infty} f(x) \, dx = 1, \text{ if } -\infty \leq X \leq \infty \]

then the function \( f(x) \) is called the probability density function (or in brief p.d.f.) of the continuous random variable \( X \).

The continuous curve \( Y = f(x) \) is called the Probability density curve (or in brief probability curve). The length of infinitesimal interval \( (x - \frac{1}{2}dx, x + \frac{1}{2}dx) \) is \( dx \) and its mid point is \( x \).

Remarks:

(1) If the range of \( X \) be finite, then also it can be expressed as infinite range.
For example,

\[ f(x) = 0, \quad \text{for } x < a \]
\[ f(x) = \phi(x), \quad \text{for } a \leq X \leq b \]
\[ f(x) = 0 \quad \text{for } x > b. \]

(2) The probability that a value of continuous variable \( X \) lies within the interval \( (c, d) \) is given by

\[ P(c \leq X \leq d) = \int_{c}^{d} f(x) \, dx \]

(3) The continuous variable always takes values within a given interval howsoever small the interval may be.

(4) If \( X \) be a continuous random variable, then

\[ P(X=k) = 0 \]

where \( k \) is a constant quantity.
2.12 CONTINUOUS PROBABILITY DISTRIBUTION:

The probability distribution of continuous random variate is called the *continuous probability distribution* and it is expressed in terms of probability density function.

2.13 CUMULATIVE DISTRIBUTION FUNCTION:

The probability that the value of a random variate X is ‘x or less than ’ is called the *cumulative distribution functions* of X and is usually denoted by F(x). In symbolic notation, the cumulative distribution function of discrete random variate X is given by

\[ F(x) = P(X \leq x) = \sum_{x \leq x_i} p(x_i) \]

The cumulative distribution function of a continuous random variate is given by

\[ F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx \]
2.14 MEAN, MEDIAN MODE AND MOMENTS FOR A CONTINUOUS DISTRIBUTION FUNCTION

(1) Mean \( \bar{M} = X = E(X) \)
\[ = \sigma \text{ if } -\infty < x < \infty \]

Also mean = \( 6 \left[ \frac{1}{2} x^2 - \frac{1}{3} x^3 \right]_0^{M_d} = \frac{1}{2} M_d \) if \( a < x < b \)

(2) Geometric Mean. If \( G \) is geometric mean, then
\[ \log G = \int_{-\infty}^{\infty} \log xf(x)dx = E(\log x) \text{ if } -\infty < x < \infty \]
\[ = \int_{a}^{b} \log xf(x)dx = E(\log x) \text{ if } a < x < b \]

(3) Harmonic Mean. Let \( H \) be the harmonic mean, then
\[ \frac{1}{H} = \int_{-\infty}^{\infty} \frac{1}{x} f(x)dx = E \left( \frac{1}{x} \right) \text{ if } -\infty < x < \infty \]
\[ \frac{1}{H} = \int_{a}^{b} \frac{1}{x} f(x)dx = E \left( \frac{1}{x} \right) \text{ if } a < x < b \]

(4) Median. The median \( M_d \) is given by
\[ \int_{-\infty}^{M_d} f(x)dx = \int_{M_d}^{\infty} f(x)dx = \frac{1}{2} \]
\[ \int_{-\infty}^{M_d} f(x)dx = \frac{1}{4} \text{ and } \int_{M_d}^{\infty} f(x)dx = \frac{1}{4} \]

(5) The lower quartile \( Q_1 \) and upper quartile \( Q_3 \) are given by
\[ \int_{-\infty}^{Q_1} f(x)dx = \frac{1}{4} \text{ and } \int_{Q_3}^{\infty} f(x)dx = \frac{1}{4} \]
\[ \int_{-\infty}^{a} f(x)dx = \frac{3}{4} \quad \text{and} \quad \int_{a}^{\infty} f(x)dx = \frac{3}{4} \]

(6) **Mode.** The mode is the value of the variant for which
\[ \frac{d}{dx} f(x) = 0 \quad \text{and} \quad \frac{d^2 f(x)}{dx^2} < 0 \]

[In other words, the mode is the value of the variant for which probability \( f(x) \) is maximum]. The condition for which is that the values obtained from \( (d/dx)f(x)=0 \) lies within the given range of \( x \).

(7) **Moment.** The \( r \)th moment about any given arbitrary value \( A \) is given by:
\[ \int_{-\infty}^{\infty} (x-m)^r f(x)dx, \quad \text{if} \quad -\infty < x < \infty \]
\[ \mu'_x = \int_{a}^{b} (x-A)^r f(x)dx, \quad \text{if} \quad a < x < b \]

(8) The Mean deviation about the mean \( m \). It is given by
\[ \int_{-\infty}^{\infty} |x-m| f(x)dx, \quad \text{if} \quad -\infty < x < \infty \]
\[ = \int_{a}^{b} |x-m| f(x)dx, \quad \text{if} \quad a < x < b \]

(9) **Variance.** For a continuous distribution the variance \( \sigma^2 \) is given by
\[ \int_{-\infty}^{\infty} (x-m)^2 f(x)dx, \quad \text{if} \quad -\infty < x < \infty \]
\[ = \int_{a}^{b} (x-m)^2 f(x)dx, \quad \text{if} \quad a < x < b \]

(10) **Standard deviation (S.D.).** The positive square root of variance is called S.D. and is denoted by \( \sigma \).
Example: For the distribution \( dF = 6(x - x^2)dx, 0 \leq x \leq 1 \), find arithmetic mean, harmonic mean, median and mode. Is it a symmetrical distribution?

Solution. For the given distribution we have

\[
\int_0^1 6(x - x^2)dx = \left[ 6 \left( \frac{1}{2} x^2 - \frac{1}{3} x^3 \right) \right]_0^1 = 1
\]

\[ f(x) = 6(x - x^2) \text{ is a p.d.f.} \]

(1) Arithmetic Mean.

\[
M = \int_0^1 x \cdot f(x)dx
\]

\[
M = \int_0^1 x \cdot 6(x - x^2)dx = 6\int_0^1 (x^2 - x^3)dx
\]

\[
= 6 \left[ \frac{1}{3} x^3 - \frac{1}{4} x^4 \right]_0^1 = 6 \left( \frac{1}{3} - \frac{1}{4} \right) = \frac{1}{2}
\]

(2) Harmonic Mean. It is given by

\[
\frac{1}{H} = \int_0^1 \frac{1}{x} \cdot 6(x - x^2)dx
\]

\[
\frac{1}{H} = 6 \left[ x - \frac{x^2}{2} \right]_0^1 = 3
\]

\[ H = \frac{1}{3} \]

(3) Median. \( M_d \) is given by:

\[
\int_0^{M_d} f(x)dx = \frac{1}{2}
\]
\[
\int_0^{M_d} 6(x-x^2) dx = \frac{1}{2}
\]

\[
6 \left[ \frac{1}{2} x^2 - \frac{1}{3} x^3 \right]_0^{M_d} = \frac{1}{2}
\]

\[
4M_d^3 - 6M_d^2 + 1 = 0
\]

\[
(2M_d - 1)(2M_d^2 - 2M_d + 1) = 0
\]

\[
M_d = \frac{1}{2} \quad \text{or} \quad M_d = \frac{1}{2}(1 \pm \sqrt{3})
\]

since \(M_d\) lies between 0 and 1.

\[
M_d = \frac{1}{2}
\]

(4) Mode. It is that value of variate \(x\) for which

\[
dy/dx = 0 \quad \text{and} \quad d^2y/dx^2 < 0
\]

\[
y = 6(x-x^2).
\]

\[
dy/dx = 6(1-2x) \quad \text{and} \quad d^2y/dx^2 = -12
\]

\[
dy/dx = 0 \Rightarrow 6(1-2x) = 0 \Rightarrow x = \frac{1}{2}
\]

Since mean, median and mode coincide, therefore the distribution is symmetrical.

---

2.15 MATHEMATICAL EXPECTATION

Definition. Let \(x\) be the discrete random variable and let its frequency distribution be as follos :

Variate : \(x_1 \quad x_2 \quad x_3 \quad \ldots \quad x_n\)
Probability : \( p_1 \quad p_2 \quad p_3 \quad \ldots \quad p_n \)

where \( \therefore p_1 + p_2 + p_3 + \ldots + p_n = \sum p = 1 \), then the mathematical expectation of \( x \) (or simply expectation of \( x \)) is denoted by \( E(x) \) and is defined by

\[
E(x) = x_1 p_1 + x_2 p_2 + x_3 p_3 + \ldots + x_n p_n = \sum_{i=1}^{n} p_ix_i = \sum p_x
\]

If \( \Phi(x) \) is a probability density function \textit{corresponding to the variate} \( x \), then

\[
E(x) = \int_{-\infty}^{\infty} \Phi(x)dx
\]

If \( \Phi(x) \) is such function of \( x \) that it takes values
\( P(A_1)(P(B/A_1) + P(A_2)(P(B/A_2) \) \textit{when} \( x \) \textit{takes the values of} \( x_1, x_2, \ldots \)

Let \( p_1, p_2, \ldots \) be their respective probabilities, then the mathematical expectation of \( \Phi(x) \), denoted by \( E[\Phi(x)] \), is defined as

\[
E[\Phi(x)] = p_1\Phi(x)_1 + p_2\Phi(x)_2 + \ldots + p_n\Phi(x)_n
\]

where \( \sum p=1 \).

Variance

The relation \( \mu_i = E[(x-E(x))^2] = E(x^2) - E(x)^2 \) denoted by \( \text{Var}(x) \), \( V(x) \) is defined as the variance of the distribution of \( x \). \( \mu \) and \( \mu' \) have the same relation as for the frequency distribution. But here the expected value of the deviation of the variate \( x \) from its mean vanishes \textit{i.e.,}

\[
E[x-E(x)] = 0.
\]
Mathematical Expectation for Continuous Random Variable

Suppose X is a continuous random variable with probability density function \( f(x) \), then mathematical expectation \( E(x) \) of \( x \) with certain restrictions is given by

\[
E(x) = \int_{-\infty}^{\infty} xf(x) \, dx
\]

Expectation of a Sum

Theorem. The expectation of the sum of two variates is equal to the sum of their expectations, i.e., if \( x \) and \( y \) are two variates then

\[
E(x + y) = E(x) + E(y)
\]

Product of Expectations

Theorem. The expectation of the product of two independent variates is equal to the product of their expectations

\[
E(xy) = E(x).E(y)
\]

2.16 COVARIANCE

Definition. Let \( x \) and \( y \) be two random variables and \( \bar{x} \) and \( \bar{y} \) be their expected values (or means) respectively. The co-variance between \( x \) and \( y \), denoted by \( \text{cov}(x,y) \), is defined as

\[
\text{cov}(x, y) = E[(x - \bar{x})(y - \bar{y})].
\]

Example 1. What is the expected value of the number of points that will be obtained in a single throw with an ordinary die? Find variance also.

Solution. The variate i.e., number showing on a die assumes the values \( 1,2,3,4,5,6 \) and probability in each case is \( 1/6 \).
given probability distribution is as follows:

\[ x: \ 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \]

\[ p: \ 1/6 \quad 1/6 \quad 1/6 \quad 1/6 \quad 1/6 \quad 1/6 \]

\[ E(x) = \sum_{i=1}^{6} p_i x_i \]

\[ = p_1 x_1 + p_2 x_2 + p_3 x_3 + \ldots + p_6 x_6 \]

\[ = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 \]

\[ = \frac{1}{6} \cdot [1 + 2 + 3 + 4 + 5 + 6] = \frac{1}{6} \cdot 21 \]

Also \[ \text{var}(x) = E(x^2) - [E(x)]^2 \]

\[ = \frac{1}{6} \cdot [1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2] - (7/2)^2 = 35/82. \]

Example 2. Find \[ \frac{\sum_{i=1}^{n} P(A_i)(P(B/A_i)P(C/A_iB))}{\sum_{i=1}^{n} P(A_i)(P(B/A_i))} \]

for the following probability distribution:

\[ x: \ 8 \quad 12 \quad 16 \quad 20 \quad 24 \]

\[ p(x): \ 1/8 \quad 1/6 \quad 3/8 \quad 1/4 \quad 1/12 \]

Solution. The mean of probability distribution is

\[ E(x) = \sum x . p(x) \]

\[ = 8 \cdot \frac{1}{8} + 12 \cdot \frac{1}{6} + 16 \cdot \frac{3}{8} + 20 \cdot \frac{1}{4} + 24 \cdot \frac{1}{12} = 16 \]
\[ E(x^2) = \sum x^2 p(x) \]

\[ = 8^2 \cdot \frac{1}{8} + 12^2 \cdot \frac{1}{6} + 16^2 \cdot \frac{3}{8} + 20^2 \cdot \frac{1}{4} + 24^2 \cdot \frac{1}{12} = 276 \]

\[ = \text{The second moment about the origin} \]

\[ E[(x-x^2)] = E(x-x^2)p(x) \]

\[ = (8-16)^2 \cdot \frac{1}{8} + (12-16)^2 \cdot \frac{1}{6} + (16-16)^2 \cdot \frac{3}{8} + (20-16)^2 \cdot \frac{1}{4} + (24-16)^2 \cdot \frac{1}{12} = 20 \]

\[ = \text{variance of the distribution.} \]

---

2.17 LET US SUM UP

**After going through this unit, you would achieved the objectives stated earlier in the unit. Let us recall what we have discussed so far –**

- Any probabilistic situation is called a random experiment.
- A sample space of a random experiment is the set of all possible outcomes of that experiment and is denoted by S.
- Of all the possible outcomes in the sample space of an experiment some outcomes satisfy a specified description, it is called an event.
- Two events are considered equally likely if one of them cannot be expected in preference to the other.
- Events are called mutually exclusive if the happening of one of them prevents or precludes the happening of the other events.
- If \( E_1 \) and \( E_2 \) are any two events then
  \[ P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2). \]
• When the happening of an event \( E_1 \) depends upon the happening of an another event \( E_2 \) then the probability of \( E_1 \) is called **conditional probability** and is denoted by \( P(E_1 | E_2) \).

• If \( E_1 \) and \( E_2 \) are two events, the respective probabilities of which are known, then the probability that both will happen simultaneously is the product of probability of \( E_1 \) and the conditional probability of \( E_2 \) when \( E_1 \) has already occurred i.e.,

\[
P(E_1 \cap E_2) = P(E_1)P(E_2 | E_1)
\]

• If the occurrence of \( E_1 \) does not depend on the occurrence of \( E_2 \) and the occurrence of \( E_2 \) does not depend on the occurrence of \( E_1 \) then \( E_1 \) and \( E_2 \) are called **independent events** otherwise they are called **dependent events**.

• An event \( B \) can be explained by a set of exhaustive and mutually exclusive hypothesis \( A_1, A_2, \ldots, A_n \). Given ‘a priori’ probabilities \( P(A_1), P(A_2), \ldots, P(A_n) \) corresponding to a total absence of knowledge regarding the occurrence of \( B \) and conditional probabilities \( P(B | A_1), P(B | A_2), \ldots, P(B | A_n) \)

the ‘a posterior’ probability \( P(A_j | B) \) of some event \( A_j \) is given by

\[
P(A_j | B) = \frac{P(A_j)P(B | A_j)}{\sum_{i=1}^{n} P(A_i)P(B | A_i)}
\]

the probability \( P(C | B) \) of materialization of another event \( C \) is given by
\[ P(A_j / B) = \sum_{i=1}^{n} P(A_i)P(B / A_i)P(C / A_iB) \]
\[ \sum_{i=1}^{n} P(A_i)P(B / A_i) \]

- A real valued function defined on a sample-space is called a random-variable. A random variable can assume only a set of real values and the values which the variable takes depends on the chance. Random variable is also called stochastic variable or simply a variate.
- When we deal with variates like weights and temperature then we know that these variates can take an infinite number of values in a given interval. Such type of variates are known as continuous variates.
- Let X be a continuous random variable and let the probability of X falling in the infinitesimal interval \((x - \frac{1}{2}dx, x + \frac{1}{2}dx)\) be expressed by \(f(x)dx\) i.e.

\[ P(x - \frac{1}{2}dx < X < x + \frac{1}{2}dx) = f(x)dx, \]

where \(f(x)\) is a continuous function of X and satisfies the following two conditions:

(i) \( f(x) \geq 0 \)

(ii) \[ \int_{a}^{b} f(x)dx = 1, \text{ if } a \leq X \leq b \]

\[ \int_{-\infty}^{\infty} f(x)dx = 1, \text{ if } -\infty \leq X \leq \infty \]

then the function \(f(x)\) is called the probability density function

- The probability distribution of continuous random variate is called the continuous probability distribution and it is expressed in terms of probability density function.

- The probability that the value of a random variate X is ‘x or less than ’ is called the cumulative distribution functions of X
Let \( x \) be the discrete random variable and let its frequency distribution be as follows:

<table>
<thead>
<tr>
<th>Variate</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>.....</th>
<th>( x_n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>( p_1 )</td>
<td>( p_2 )</td>
<td>( p_3 )</td>
<td>.....</td>
<td>( p_n )</td>
</tr>
</tbody>
</table>

where \( \sum p = 1 \), then the mathematical expectation of \( x \) (or simply expectation of \( x \)) is denoted by \( E(x) \) and is defined by

\[
E(x) = p_1x_1 + p_2x_2 + p_3x_3 + \ldots + p_nx_n = \sum_{i=1}^{n} p_i x_i = \sum p x
\]

- The expectation of the sum of two variates is equal to the sum of their expectations, i.e., if \( x \) and \( y \) are two variates then

\[
E(x + y) = E(x) + E(y)
\]

- The expectation of the product of two independent variates is equal to the product of their expectations

\[
E(xy) = E(x).E(y)
\]

- Let \( x \) and \( y \) be two random variables and \( \bar{x} \) and \( \bar{y} \) be their expected values (or means) respectively. The co-variance between \( x \) and \( y \), denoted by \( \text{cov}(x,y) \), is defined as

\[
\text{cov}(x, y) = E[(x - \bar{x})(y - \bar{y})].
\]
Exercise

Q.1) A coin is tossed twice. What is the probability that at least one head occurs.

Q.2) A card is drawn from an ordinary deck. Find the probability that it is a heart.

Q.3) What is the probability of getting a total of 7 or 11 when a pair of dice is tossed?

Q.4) A coin is tossed 6 times in succession. When is the probability that at least one head occurs?

Q.5) A card is drawn from an ordinary deck and we are told that it is red. What is the probability that the card is greater than 2 but less than 9.

Q.6) Find a formula for the probability distribution of the random variable X representing the outcome when a single die is rolled once.
UNIT – 3. THEORETICAL DISTRIBUTIONS

Structure

3.1. Introduction

3.2. Objectives

3.3. Theoretical Distributions

3.4. Binomial Distribution
   3.4.1. Constants of the Binomial Distribution
   3.4.2. Renovsky Formula:
   (Recurrence relation for the moment of Binomial distribution).
   3.4.3. Moment Generating Function of Binomial Distribution
   3.4.4. Mode of Binomial Distribution
   3.4.5. Illustrative Examples

3.5. Poisson’s Distribution
   3.5.1. Constants of the Poisson’s Distribution
   3.5.2. Recurrence relation for the moment of Poisson’s distribution.
   3.5.3. Moment Generating Function of Poisson’s Distribution
   3.5.4. Mode of Poisson’s Distribution
   3.5.5. Illustrative Examples

3.6. Normal Distribution

3.6.2. Moment Generating Function of Normal Distribution

3.6.3. Illustrative Examples

3.7. Rectangular Distribution

3.7.1. Constants of the Rectangular Distribution

3.7.2. Moment Generating Function of Rectangular Distribution

3.7.3. Illustrative Examples

3.8. Exponential Distribution

3.8.1. Constants of the Exponential Distribution

3.8.2. Moment Generating Function of Exponential Distribution

3.8.3. Illustrative Examples

3.9. Summary of Unit

3.10. Assignment

3.11. Check your progress

3.12. Point of Discussion

3.13. Suggested Study Material

3.1. Introduction: We are already familiar to the concepts of frequency distribution, measure of central tendency measure of dispersion mean deviation, standard deviation, moment, skewness, kurtosis, theory of probability, mathematical expectation and moment generating functions. In this unit we shall confine over selves to the study of theoretical distribution – Binomial, Poisson, Normal, Rectangular and Exponential distributions. Further we deal with the properties and applications of these distributions. These
distributions are divided into two parts – Discrete Theoretical Distributions and Continuous Theoretical Distributions. Binomial and Poisson distributions are Discrete Theoretical Distributions while Normal, Rectangular and Exponential are Continuous Theoretical Distributions. If certain hypothesis is assumed, it is sometimes possible to derive mathematically what the frequency distributions of certain universes should be. Such distributions are called ‘Theoretical Distributions’. The Binomial Distribution was discovered by James Bernoulli in 1700 and therefore it is also called the Bernoulli Distribution. The Poisson distribution is a particular limiting form of the Binomial distribution; it was first discovered by a French Mathematician S.D. Poisson in 1837. In 1733 Demoivre made the discovery of Normal (or Gaussian) distribution as a limiting form of Binomial distribution, it is defined by the probability density function. Rectangular (or Uniform) and Exponentials distributions are also defined by different probability density functions. These distributions serve as the guiding instrument in researchers in the physical, social sciences and in medicine, agriculture and engineering. These are indispensable tool for the analysis and the interpretation of the basic data obtained by observation Experiment.

3.2: Objectives: After the end of the unit the student will be able to understand/know the

1. Concepts of Theoretical Distributions
2. Difference between the Discrete and Continuous Theoretical distributions
3. Binomial and Poisson Distributions – Properties, constants and applications
4. Normal, Rectangular and Exponential Distributions – Properties, constants and applications
5. Able to solve the problems based on above distributions
6. Able to use of these distributions in every day life calculations
7. Merits and De-merits of these distributions

3.3. Theoretical Distributions: When the frequency distributions are made by collecting the data in direct form. Then such distributions are called observed frequency distributions.

When the frequency distributions are made by obtaining probable (or expected) frequency by using mathematical methods on the basis of definite hypothesis (or assumptions) then such distributions are said to be theoretical frequency distributions. Theoretical distributions are obtained by probability distributions. If the probabilities are
assumed to the relative frequencies then the probability distributions are said to be theoretical frequency distributions. Theoretical distributions are divided into two following parts:

1. Discrete theoretical frequency distributions
   (a) Binomial distribution
   (b) Poisson distribution

2. Continuous theoretical frequency distributions
   (a) Normal distribution
   (b) Rectangular distribution
   (c) Exponential distribution

Other synonyms of theoretical frequency distributions are mathematical frequency distributions, ideal distributions, and expected frequency distributions.

We can understand the theoretical distribution by following example:

Example: If we throw four coins 80 times then by the theory of probability the theoretical distribution is given as:

<table>
<thead>
<tr>
<th>No of tails</th>
<th>Probability</th>
<th>Expected frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$\frac{1}{16}$</td>
<td>$80 \times \frac{1}{16} = 5$</td>
</tr>
<tr>
<td>1</td>
<td>$\frac{4}{16}$</td>
<td>$80 \times \frac{4}{16} = 20$</td>
</tr>
<tr>
<td>2</td>
<td>$\frac{6}{16}$</td>
<td>$80 \times \frac{6}{16} = 30$</td>
</tr>
<tr>
<td>3</td>
<td>$\frac{4}{16}$</td>
<td>$80 \times \frac{4}{16} = 20$</td>
</tr>
<tr>
<td>4</td>
<td>$\frac{1}{16}$</td>
<td>$80 \times \frac{1}{16} = 5$</td>
</tr>
<tr>
<td>Total</td>
<td>1</td>
<td>80</td>
</tr>
</tbody>
</table>
3.4. Binomial Distribution: Let there be an event, the probability of its being success is p and the probability of its failure is q in one trial, so that p + q = 1. Let the event be tried n times and suppose that the trials are

(i) Independent

(ii) Number of trials are finite and

(iii) The probability p of success is the same in every trial.

The number of successes in n trials may be 0, 1, 2, ..., r, ..., n.

The probability that the first r trials are successes and the remaining n – r are failures is \( p^r q^{n-r} \). But we are to consider all the cases where any r trials are successes, since out of n, r can be chosen \( ^nC_r \) ways, then the probability \( p(r) \) [or \( b(n, p, r) \) of r successes out of n independent trials is given by

\[
p(r) = \binom{n}{r} p^r q^{n-r}.
\]

Thus the number success, can take the values 0, 1, 2, ..., r, ..., n with corresponding probabilities \( q^n, \binom{n}{1} p q^{n-1}, \binom{n}{2} p^2 q^{n-2}, ..., \binom{n}{r} p^r q^{n-r}, ..., p^n \). The probability of the number of successes so obtained is called the binomial probability distribution for the obvious reason that the probabilities are the varies terms in the binomial expansion of \( (q + p)^n \). The sum of the probabilities

\[
= \sum_{r=0}^{n} b(n, p, r) = \sum_{r=0}^{n} \binom{n}{r} p^r q^{n-r} = (q + p)^n = 1.
\]

Definition: The probability distribution of random variable \( x \) is called Binomial distribution if only take non-negative values and its distribution is given by,

\[
P(r) = p(X = r) = \begin{cases} \binom{n}{r} p^r q^{n-r}, & r = 0, 1, 2, ..., n; \\ 0, & r \neq 0, 1, 2, ..., n \end{cases}
\]

The binomial distribution contains two independent constants viz, n and p (or q), these are called the parameter of the binomial distribution. If \( p = q = \frac{1}{2} \), the binomial distribution is called symmetric and when \( p \neq q \) it is called skew symmetric distribution.

Let the \( n \) independent trials constitute one experiment and let this experiment repeated \( N \) times where \( N \) is very large. In these \( N \) sets there will be few sets in which there is no success, a few sets of one success, a few sets of two successes, and so on. Hence in all the \( N \) sets, the number of sets with r successes is \( N \binom{n}{r} p^r q^{n-r} \). Therefore the number of sets corresponding to the number of successes 0, 1, 2, ..., r, ..., n, are respectively \( N q^n, \binom{n}{1} p q^{n-1}, \binom{n}{2} p^2 q^{n-2}, ..., \binom{n}{r} p^r q^{n-r}, ..., N p^n \). Hence for \( N \) sets of \( n \) trials the Theoretical frequency distribution or Expected frequencies distribution of 0, 1, 2, ..., r, ..., n successes are given by the successive terms in the expression
\[ N[ q^n + {n\choose 1} p q^{n-1} + {n\choose 2} p^2 q^{n-2} + \ldots + {n\choose r} p^r q^{n-r} + \ldots + p^n ] \]

which is the binomial expansion of \( N(q + p)^n \). This is called the Binomial Theoretical Frequency Distribution or simply Binomial Distribution.

3.4.1. Constants of the Binomial Distribution:

Moments about origin: \( \mu_k' = E(X^k) = \sum_{r=0}^{n} r^k \cdot p(r) = \sum_{r=0}^{n} r^k \cdot {n\choose r} p^r q^{n-r}, \quad q + p = 1, \)

\((k = 1, 2, 3, 4, \ldots)\).

First moment

\[ \mu_1' = E(X) = \sum_{r=0}^{n} r \cdot {n\choose r} p^r q^{n-r}, \quad q + p = 1. \]

\[ = \sum_{r=1}^{n} r \cdot {n\choose r-1} p^{r-1} q^{n-r} = np \sum_{r=1}^{n} {r-1\choose r-1} p^{r-1} q^{(n-1)-(r-1)} \]

\[ = np[q^{n-1} + {n-1\choose 1} q^{n-2} p + {n-1\choose 2} q^{n-3} p^2 + \ldots + {n-1\choose n-1} q^1 p^{n-1} ] \]

\[ = np(q + p)^{n-1} = np. \quad \text{(Since } q + p = 1) \]

Hence  \( \mu_1' = \text{Mean} = np. \)

Second moment

\[ \mu_2' = E(X^2) = \sum_{r=0}^{n} r^2 \cdot {n\choose r} p^r q^{n-r}, \quad q + p = 1. \]

\[ = \sum_{r=0}^{n} \left( (r-1)^2 + r \right) \cdot {n\choose r} p^r q^{n-r} \]

\[ = \sum_{r=0}^{n} r(r-1) \cdot {n\choose r} p^r q^{n-r} + \sum_{r=0}^{n} {n\choose r} p^r q^{n-r} \]

\[ = n (n-1)p^2 \sum_{r=2}^{n} \cdot {n-2\choose r-2} p^{r-2} q^{(n-2)-(r-2)} + np \]

\[ = n (n-1)p^2(q + p)^{n-2} + np \quad \text{(since } q + p = 1) \]

\[ = np\{np + (1 - p)\} = np\{np + q\} = npq + n^2 p^2. \]
Similarly third and fourth moment about origin are

\[
\mu_3' = E(X^3) = \sum_{r=0}^{n} r^3 \cdot {n \choose r} p^r q^{n-r} = \sum_{r=0}^{n} \left\{ r(r-1)(r-2) + 3r(r-1) + r \right\} \cdot {n \choose r} p^r q^{n-r},
\]

\[
= n(n-1)(n-2) p^3 + 3n(n-1)p^2 + np \quad \text{and}
\]

\[
\mu_4' = E(X^4) = \sum_{r=0}^{n} r^4 \cdot {n \choose r} p^r q^{n-r},
\]

\[
= \sum_{r=0}^{n} \left\{ r(r-1)(r-2)(r-3) + 6r(r-1)(r-2) + 7r(r-1) + r \right\} \cdot {n \choose r} p^r q^{n-r},
\]

\[
= n(n-1)(n-2)(n-3) p^4 + 6n(n-1)(n-2)p^3 + 7n(n-1)p^2 + np
\]

Moments about mean: \( \mu_n \quad (n = 1, 2, 3, 4, \ldots) \)

\[
\mu_1 = 0 \quad \text{always,} \quad \mu_2 = \mu_2' - (\mu_1')^2 = npq + n^2p^2 - n^2p^2 = npq \quad \text{or, variance } \sigma^2 = npq \quad \text{and standard deviation (S.D.)} = \sqrt{npq},
\]

\[
\mu_3 = \mu_3' - 3 \mu_2'\mu_1 + 2(\mu_1')^3
\]

\[
= n(n-1)(n-2) p^3 + 3n(n-1)p^2 + np - 3 \{ npq + n^2p^2 \} np + 2 \{ np \} p^3,
\]

\[
= np\{ (n-1)(n-2) p^2 + 3(n-1)p + 1 - 3 npq - 3 n^2 p^2 + 2 \{ np \} p^2 \},
\]

\[
= np\{ n^2 p^2 - 3np^2 + 2p^2 + 3np - 3p + 1 - 3 np - n^2 p^2 \},
\]

\[
= np\{ - 3np^2 + 2p^2 + 3np - 3p + 1 \} = np(1-p) \{ 1-2p \},
\]

\[
= npq \{ 1 - 2(1-q) \} = npq(2q-1) = npq(q-p),
\]

\[
\mu_4 = \mu_4' - 4 \mu_3'\mu_1 + 6 \mu_2'\mu_1 - 3 \mu_1'^4
\]

\[
= n(n-1)(n-2)(n-3) p^4 + 6n(n-1)(n-2)p^3 + 7n(n-1)p^2 + np
\]

\[- 4[n(n-1)(n-2) p^3 + 3n(n-1)p^2 + np]np + 6 \{ n(n-1)p^2 + np \} n^2 p^2 - 3 \{ np \} p^4, \]

\[
= p^4 \{ n(n-1)(n-2)(n-3) - 4n^2 (n-1)(n-2) + 6n^3(n-1) - 3n^4 \}
\]

\[
+ p^3 \{ 6n(n-1)(n-2) - 12 n^2(n-1) + 6n^2 \} + p^2 \{ 6n(n-1) - 4 n^2 \} + np,
\]

\[
= p^4 \{ 3n^2 - 6n \} + p^3 \{ 12 n - 6n^2 \} + p^2 \{ 3n^2 - 7n \} + np,
\]

\[
= 3n^2 \{ p^4 - 2p^3 + p^2 \} - n(6p^4 - 12p^3 + 6p^2) - np^2 + np,
\]
\[= 3n^2 p^2 (p^2 - 2p + 1) - 6np^2 (p^2 - 2p + p^2) + np(1 - p),\]
\[= 3n^2 p^2 (1 - p)^2 - 6np^2 (1 - p)^2 + np(1 - p),\]
\[= 3n^2 \frac{q^2}{p^2} - 6np^2 \frac{q^2}{p^2} + npq = npq[1 + 3(n - 2)pq].\]

Karl Pearson’s Coefficients: \(\beta_1, \beta_2, \gamma_1, \gamma_2\)

\[
\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(npq(q - p))^2}{(npq)^3} = \frac{(1 - 2p)^2}{npq},
\]
\[
\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3n^2 p^2 q^2 + npq(1 - 6pq)}{(npq)^2} = 3 + \frac{(1 - 6pq)}{npq},
\]
\[
\gamma_1 = \sqrt{\beta_1} = \frac{1 - 2p}{\sqrt{npq}}, \quad \gamma_2 = \beta_2 - 3 = \frac{(1 - 6pq)}{npq}.
\]

3.4.2. Renovsky Formula:

(Recurrence relation for the moments of Binomial Distribution). For Binomial distribution \((q + p)^n\),

\[
\mu_{k+1} = np \left( nk \mu_k + \frac{d\mu_k}{dp} \right), \text{ where } \mu_k \text{ is the } k^{th} \text{ moment about mean.}
\]

Proof: Since \(\mu_k = \sum_{r=0}^{n} (r - \mu_1)^k \cdot p(r) = \sum_{r=0}^{n} (r - np)^k \cdot C_r \cdot p^r \cdot q^{n-r}, \ q + p = 1,\)

On differentiating with respect to \(p, \) we get

\[
\frac{d\mu_k}{dp} = \sum_{r=0}^{n} (-1)^k n (r - np)^{k-1} \cdot C_r \cdot p^r (1 - p)^{n-r} + \sum_{r=0}^{n} ((r - np)^k \cdot C_r \cdot r p^{r-1} (1 - p)^{n-r} - (n - r) p^r (1 - p)^{n-r-1})
\]
\[
= -nk \mu_k + \frac{1}{pq} \sum_{r=1}^{n} C_r \cdot p^r \cdot q^{n-r} \cdot (r - np)^k (rq - np + rp),
\]

Therefore \(np \left( nk \mu_k + \frac{d\mu_k}{dp} \right) = \sum_{r=0}^{n} C_r \cdot p^r \cdot q^{n-r} \cdot (r - np)^{k+1} = \mu_{k+1}.\)

If we put \(k = 1, 2, 3, \ldots \) we have
\[\mu_2 = pq \left(n\mu_0 + \frac{d\mu_1}{dp}\right) = pq(0 + n) = npq, \sin ce \mu_0 = 1 \text{ and } \mu_1 = 0,\]
\[\mu_3 = pq \left(n^2 \mu_1 + \frac{d\mu_2}{dp}\right) = pq(nq - np + 2n0) = npq(q - p),\]
\[\mu_4 = pq \left(n^3 \mu_2 + \frac{d\mu_3}{dp}\right) = pq[3nnp(1 - p) + n(6p^2 - 6p + 1)],\]
\[= 3(npq)^2 + npq(1 - 6pq), \text{ etc.}\]

3.4.3. Moment Generating Function of Binomial Distribution:

Let \( r \) have a Binomial Distribution with probability function
\[p(r) = \binom{n}{r} p^r q^{n-r}; \quad r = 0,1,2, ... , n, \quad \text{where } r = \text{number of successes in } n \text{ trials and } p = \text{probability of success in a single trial, } q = 1 - p.\]

Then m.g.f. about origin: \( M_0(t) = E(e^{tr}) = \sum_{r=0}^{n} e^{tr} \binom{n}{r} p^r q^{n-r} = \sum_{r=0}^{n} \binom{n}{r} (pe^t)^r q^{n-r} = (q+ pe^t)^n\)

m.g.f. about mean \( np \): \( M_{np}(t) = E[e^{t(r - np)}] = E[e^{t(r - np)}] = e^{np^t} E(e^{tr}) = e^{np^t} M_0(t) \]
\[= e^{np^t} E(e^{tr}) = e^{np^t}(q+ pe^t)^n = (qe^{pt} + pe^{qt})^n,\]
\[= (qe^{pt} + pe^{qt})^n.\]

Generation of moments of Binomial Distribution from m.g.f.: Since

\[M_0(t) = (q+ pe^t)^n = \left[ q + p\left( 1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + ... \right) \right]^n \]
\[= \left[ 1 + pt + \frac{pt^2}{2!} + \frac{pt^3}{3!} + ... \right]^n \]
\[= 1 + n C_1 \left[ pt + \frac{pt^2}{2!} + \frac{pt^3}{3!} + ... \right]^n + n C_2 \left[ pt + \frac{pt^2}{2!} + \frac{pt^3}{3!} + ... \right]^n + ... \quad \text{(i)}\]

From (i), we get \( \mu_1' = \text{coefficient of } t = np \),
\[\mu_2' = \text{coefficient of } t^2/2! = np + n C_2 \left( p^2 = np + \frac{n(n-1)}{2} .2p^2 \right).\]
\[= np + n(n - 1)p^2 = np + n^2 p^2 - np = np(1 - p) + n^2 p^2.\]

Also, \(\mu_2 = \mu_2' - (\mu_1')^2 = np + (np)^2 - (np)^2 = \text{variance and S.D. } \sigma = \sqrt{npq}.

Similarly \(\mu_3', \mu_4', \ldots \) and \(\mu_3, \mu_4, \ldots \) etc may be obtained.

Alternative Method: \(\mu_r' = \left[\frac{d'}{dt'}[M_0(t)]_{t=0}\right] = \left[\frac{d'}{dt'}[(q + pe')^n]\right]_{t=0}

On putting \(r = 1, 2, 3, \ldots \), we get

\[
\begin{align*}
\mu_1' &= \left[\frac{d}{dt}[q + pe']\right]_{t=0} = npq = np + (np)^2 = \frac{np}{n - 1}.
\end{align*}
\]

\[
\begin{align*}
\mu_2' &= \left[\frac{d^2}{dt^2}[q + pe']\right]_{t=0} = npq^2 = np + (np)^2 - npq.
\end{align*}
\]

3.4.4. Mode of Binomial Distribution: (The most probable number of success in a series of \(n\) independent trials; the probability of success in each trials being \(p\)).

Here we are to find the number of successes which has a greater probability than any other. Let the probability of \(r\) successes is greater than or equal to that of \(r - 1\) or \(r + 1\) successes, i.e., \(p(r - 1) \leq p(r) \geq p(r + 1).\)

If \(\binom{n}{r-1}p^{r-1}q^{n-r+1} \leq \binom{n}{r}p^rq^{n-r} \geq \binom{n}{r+1}p^{r+1}q^{n-r+1}\),

Simplifying we get

\[
\frac{r}{n-r+1} \cdot \frac{q}{p} \leq \frac{n-r}{r+1} \cdot \frac{p}{q} \implies rq \leq np - rp + p \text{ and } rq + q \geq np - rp.
\]

\[
\implies np - q \leq r \leq np + p \implies (n + 1) p - 1 \leq r \leq np + p.
\]

Case 1. If \((n + 1)p = k\) (an integer), then probability will increase till \(r = k\) and it will be the same for \(r = k - 1\) and after that it will begin to decrease.

Case 2. If \(np = \text{an integer} + \text{a fraction}\), then probability is maximum when \(r = \text{the integral part of} \ (np + p)\).

Here we note that
(i) If \( np \) is a whole number, the mean of the Binomial distribution coincides with the greatest term. Since frequency of \( r \) successes is greater than that of \( r - 1 \) successes if \( \binom{n}{r} p^r q^{n-r} < \binom{n}{r+1} p^{r+1} q^{n-r-1} \)

\[
\frac{p}{r} > \frac{q}{n-r+1} \quad \text{or} \quad r < np + p.
\]

Similarly the frequency of \( r \) successes is greater than that of \( r + 1 \) successes if

\[
\binom{n}{r+1} p^{r+1} q^{n-r-1} > \binom{n}{r} p^r q^{n-r} \quad \text{implies that} \quad r > np - q. \]

Thus if \( np \) is a whole number \( r = np \) gives the greatest term and also the mean of the Binomial.

(ii) The difference of mean and mode of the Binomial distribution is not greater than unity. Since the mean of Binomial distribution = \( np \), also there are three cases for mode are:

(a) If \( np \) is positive integer then mode and mean are equal, therefore

\[ |\text{mean} - \text{mode}| = 0 < 1. \]

(b) If \( np + p \) is positive integer then mode is \( np + p \), therefore

\[ |\text{mean} - \text{mode}| = |np + p - np| = p < 1. \]

(c) If \( np + p \) is not positive integer then mode say \( r \) is given by \( np + p - 1 \leq r \leq np + p \)

implies that \( p - 1 \leq r - np \leq p \) or \( -p \leq r - np \leq p \) or \( -p \leq \text{mode} - \text{mean} \leq p \), therefore

\[ |\text{mean} - \text{mode}| \leq 1. \]

3.4.5. Illustrative Examples.

Example 1. Criticize the statement: For any Binomial distribution mean is 5 and standard deviation is 3.

Solution. For Binomial distribution, we have given that, mean = \( np = 5 \) and standard deviation = \( \sqrt{npq} = 3 \),

\[ \Rightarrow \frac{npq}{np} = \frac{9}{5} \quad \Rightarrow q = 1.8 > 1. \]

Hence the given statement is not correct.

Example 2. The probability of a head in a single tossing of a biased coin is \( 3/5 \). Find the most probable number of heads and the mean of number of heads in 99 tossing of a coin.

Solution: Let on tossing 99 times the number of getting heads are = 0, 1, 2, 3, \ldots, 99. We have given that and here \( p = 0.6 \), \( q = 0.4 \). Therefore the probability, distribution is
\[ P(X = r) = ^nC_r p^r q^{n-r} = ^{99}C_r (.6)^r(.4)^{99-r}; \ r = 0, 1, 2, 3, \ldots, 99 \ \text{and mean} = 99 \times 0.6 = 59.4 \ (\text{not an integer}), \ \text{therefore two modes are}, \ np + p = 59.4 + 0.6 = 60 \ \text{and} \ np + p - 1 = 59, \ \text{where} \ P(X = 59) = ^{99}C_{59}(.6)^{59}(.4)^{40} \ \text{and} \ P(X = 60) = ^{99}C_{60}(.6)^{60}(.4)^{39} \ \text{are same.}

Example 3. A variate assumes values 0, 1, 2, \ldots, n whose frequencies are proportional to the binomial coefficients \(^nC_1, ^nC_2, \ldots, ^nC_n\). Show that variance is half of the mean.

Solution: Given:

<table>
<thead>
<tr>
<th>x</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>\ldots</th>
<th>n</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>1</td>
<td>(^nC_1)</td>
<td>(^nC_2)</td>
<td>\ldots</td>
<td>(^nC_n)</td>
<td>((1 + 1)^n = 2^n)</td>
</tr>
</tbody>
</table>

This distribution is a Binomial distribution in which \(p = q = 1/2\)

Now since mean = np = n \times \frac{1}{2} = n/2 and Variance = npq = n \times \frac{1}{2} \times \frac{1}{2} = n/4.

Comparing (1) and (2), we get Variance = \(\frac{1}{2}\) mean.

Example 4. If the m.g.f of a random variable \(X\) is \(\left( \frac{1}{3} + \frac{2}{3}e^t \right)^{5}\), find \(P(X = 2)\).

Solution: We have given that \(M_0(t) = (q + pe^t) = \left( \frac{1}{3} + \frac{2}{3}e^t \right)^{5}\), implies that \(p = 2/3, \ q = 1/3\) and \(n = 5\). Therefore \(P(X = 2) = ^5C_2 (2/3)^2(1/3)^3 = 40/243\).

Example 5. Assuming that half the population are consumers of chocolate, so that the chance of an individual being a consumer is \(1/2\), and assuming that 100 investigators each take 10 individuals to see whether they are consumers, how many investigators would you expect to report that three people or less were consumers?

Solution: Here, \(p = \frac{1}{2}, \ n = 10, \ N = 100\). Binomial distribution is \(100\left( \frac{1}{2} + \frac{1}{2} \right)^{10}\)

Number of investigators to report that no person is consumer, one person is consumer, 2 persons are consumers, 3 are consumers =

\[
100\left[ \left( \frac{1}{2} \right)^{10} + ^{10}C_1\left( \frac{1}{2} \right)^{10} + ^{10}C_2\left( \frac{1}{2} \right)^{10} + ^{10}C_3\left( \frac{1}{2} \right)^{10} \right]
\]

\[
= \frac{100}{2^{10}}(1 + 10 + 45 + 120) = \frac{17600}{1024} = 17 \text{ appr.}
\]

Example 6. Determine the Binomial distribution for which the mean is 4 and variance is 3 and find its mode.
Solution: Let \( P(X = r) = ^nC_r p^r q^{n-r} \), \( r = 0, 1, 2, \ldots, n \). It is given that mean = np = 4 and variance = npq = 3 implies that \( q = \frac{3}{4}, \ p = \frac{1}{4} \) and \( n = 16 \).

Thus the Binomial distribution is, \( P(X = r) = ^{16}C_r \left(\frac{1}{4}\right)^r \left(\frac{3}{4}\right)^{16-r}, \ r = 0, 1, 2, 3, \ldots, 16 \).

Now, \( np + p = 4 + \frac{1}{4} = 4.25 \) is not an integer. Hence there is one mode = integral part of np + p = 4.

Alternatively: Since np is an integer, mode = mean = np = 4.

Example 7. The following results are obtained when 100 batches of seeds were allowed to germinate on damp filter paper in a laboratory: \( \beta_1 = \frac{1}{5}, \beta_2 = \frac{89}{30} \). Determine the Binomial distribution. Calculate the expected frequency for \( x = 8 \) assuming \( p > q \).

Solution: For Binomial distribution we have given that

\[
\beta_1 = \frac{(1 - 2p)^2}{npq} = \frac{1}{15} \quad \text{and} \quad \beta_2 = 3 + \frac{1 - 6pq}{npq} = \frac{89}{30}
\]

\[
\Rightarrow \frac{(1 - 2p)^2}{npq} = \frac{1}{15} \quad \text{and} \quad \frac{1 - 6pq}{npq} = \frac{89}{30} - 3 = -\frac{1}{15}
\]

\[
\Rightarrow \frac{1 - 2p}{1 - 6pq} = \frac{1}{15} \quad \text{and} \quad \frac{1 - 6pq}{npq} = \frac{89}{30} - 3 = -\frac{1}{15}
\]

\[
\Rightarrow (1 - 2p)^2 = -2[1 - 6p(1 - p)],
\]

\[
\Rightarrow 1 - 4p + 4p^2 = -2 - 12p + 12p^2,
\]

\[
\Rightarrow 16p^2 + 16p + 3 = 0,
\]

\[
\Rightarrow (4p - 1)(4p - 3) = 0,
\]

\[
\Rightarrow p = \frac{1}{4} \quad \text{or} \quad p = \frac{3}{4} \quad \text{and so} \quad q = \frac{3}{4} \quad \text{or} \quad q = \frac{1}{4}.
\]

Since it is given that \( p > q \), hence we take \( p = \frac{3}{4} \) and \( q = \frac{1}{4} \) is valid. Further

\[
\frac{(1 - 2p)^2}{npq} = \frac{1}{15} \Rightarrow \left(1 - 2 \times \frac{3}{4}\right)^2 = \frac{1}{15} \Rightarrow n = 20. \quad \text{Thus the required Binomial Distribution}
\]

is \( (q + p)^n = \left(\frac{1}{4} + \frac{3}{4}\right)^{20} \) where \( P(X = r) = ^{20}C_r (\frac{3}{4})^r (\frac{1}{4})^{20-r} \). For \( x = 8 \), the expected frequency = \( 10 \times ^{20}C_8 (\frac{3}{4})^8 (\frac{1}{4})^{12} \).
Example 8. Is the sum of two independent Binomial variates a Binomial?

Solution: Let $x_1$ and $x_2$ be two independent Binomial variates with parameters $n_1, p_1$ and $n_2, p_2$ respectively. Then

$$M(t) \text{ of } x_1 = (q_1 + p_1 e^t)^{n_1} \quad \text{and} \quad M(t) \text{ of } x_2 = (q_2 + p_2 e^t)^{n_2}.$$  \hspace{1cm} \text{Let } x = x_1 + x_2. \quad \text{Then moment generation function of } x = (q + p e^t)^{n_1+n_2} \text{ which is not of the form } (q + pe^t)^n. \quad \text{Hence } x \text{ is not a Binomial variate. But if } p_1 = p_2 = p \text{ and } q_1 = q_2 = q, \quad \text{then moment generation function of } x = (q + p e^t)^{n_1+n_2} \text{ which is the Binomial variate with parameters } (n_1 + n_2) \text{ and } p.

Example 9. The following data are the number of seeds germinating out of 10 on damp filter for 80 sets of seeds. Fit a Binomial distribution to these data:

<table>
<thead>
<tr>
<th>$x$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f$</td>
<td>6</td>
<td>20</td>
<td>28</td>
<td>12</td>
<td>8</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>80</td>
</tr>
</tbody>
</table>

Here $n = 10, N = 80$ and total frequency = 80. Then

A.M. $= \frac{\sum fx}{\sum f} = \frac{1\times20 + 2\times28 + 3\times12 + 4\times8 + 5\times6}{80} = \frac{174}{80} = 2.175$. Now, mean = $np$ implies that $p = 0.2175$ and $q = .7825$. Hence the required Binomial distribution of the given data is $10^\text{th}$ from this expansion the successive frequencies of 0, 1, 2, ..., 10 successes are 6.9, 19.1, 24.0, 17.8, 8.6, 2.9, 0.7, 0.1, 0, 0, 0 respectively.

Example 10. If in a Binomial distribution $(q + p)^n$, $n = 5, P(X = 1) = 0.4096$ and $P(X = 1) = 0.2048$, then find $p$.

Solution: We have given that $P(X = 1) = 5C_1 p^1 q^4 = 0.4096$ and $P(X = 2) = 5C_2 p^2 q^3$ implies $\frac{P(X = 1)}{P(X = 2)} = \frac{5pq^4}{10(pq)^2} = \frac{(1-p)}{2p} = \frac{0.4096}{0.2048}$. \quad \Rightarrow \quad p = 0.2.$

Example 11. 6 dices are thrown 729 times. Find the probability of obtaining five or six at least three dices.

Solution: We know that the probability of getting five or six on a throw of a die = $2/6 = 1/3 = p$. Then $q = 1 - p = 2/3$. Therefore the probability of getting five or six on at least three dices out of six = $P(3) + P(4) + P(5) + P(6)$,

$$= 6C_3 p^3 q^3 + 6C_4 p^4 q^2 + 6C_5 p^5 q + p^6$$
$$= 6C_3 (1/3)^3 (2/3)^3 + 6C_4 (1/3)^4 (2/3)^2 + 6C_5 (1/3)^5 (2/3) + (1/3)^6$$
$$= (1/3)^6[6C_3 2^3 + 6C_4 2^2 + 6C_5 2 + 1] = 233/729 = P(say).$$
Hence the required probability $= np = 729 \times (233/729) = 233$.

3.5. Poisson’s Distribution (Limiting form of Binomial Distribution): **Poisson Distribution** is a particular limiting form of the Binomial distribution when $p$ (or $q$) is very small and $n$ is large enough such that $np$ (or $nq$) is a finite constant say $np = m$.

We know that under the conditions, $p(r)$ the probability of $r$ successes in the Binomial distribution is given by,

$$p(r) = p(X = r) = C_n^r p^r q^{n-r} = C_n^r (1 - p)^{n-r}$$

$$= \frac{n(n-1)(n-2)\ldots(n-r+1)}{r!} p^r (1 - \frac{np}{n})^{n-r}$$

$$= (1 - \frac{1}{n})(1 - \frac{2}{n})\ldots(1 - \frac{r-1}{n}) \cdot \frac{(np)^r}{r!} \cdot \frac{(1 - np/n)^n}{(1 - np/n)^r}.$$

Now the probability of $r$ success in Poisson’s distribution,

$$p(r) = b(n, p; r) = \lim_{\frac{n \rightarrow \infty}{np \rightarrow m}} \frac{(1 - \frac{1}{n})(1 - \frac{2}{n})\ldots(1 - \frac{r-1}{n})}{r!} \cdot \frac{m^r}{(1 - m/n)^r} \cdot \frac{(1 - np/n)^n}{(1 - np/n)^r}$$

$$= \frac{m^r}{r!} e^{-m}.$$

It is called the **Poisson’s distribution** with parameter $m$. Therefore the chances of $0, 1, 2, 3, \ldots, r$ successes are $e^{-m}, \frac{m}{r!} e^{-m}, \frac{m^2}{2!} e^{-m}, \ldots, \frac{m^r}{r!} e^{-m}$ respectively.

**Definition**: The probability distribution of random variable $X$ is called Poisson’s distribution if $X$ can assume non-negative integral values only and its distribution is given by, $P(r) = p(X = r) = \{ \frac{m^r}{r!} e^{-m}, r = 0, 1, 2, 3, \ldots, n \}$.

**Note**: (i) $m$ is said to be parameter of Poisson’s distribution.

(ii) This distribution is useful in solving the problems of following types (some examples) (a) The number of cars passing through a certain street in time $t$.

(b) The number of defective screws per box of 100 screws.

(c) The number of deaths in a district in one year by a rare disease.
(d) The number of suicides or deaths by heart attack in time \( t \).

(e) The number of pieces of certain merchandise sold by a store in time \( t \).

(f) The number of printing mistakes at each page of the book.

(g) The number of calls of telephone from a particular switch board.

(i) Number of accidents per unit time.

3.5.1. Constants of the Poisson Distribution:

Moments about origin: \( \mu_k' = E(X^k) = \sum_{r=0}^{\infty} r^k \cdot p(r) = \sum_{r=0}^{\infty} r^k \cdot \frac{m^r e^{-m}}{r!} \),

\( (k = 1, 2, 3, 4, \ldots) \).

First moment

\[
\mu_1' = E(X) = \sum_{r=0}^{\infty} r \cdot \frac{m^r e^{-m}}{r!} = \sum_{r=0}^{\infty} \frac{m^r e^{-m}}{(r-1)!} = e^m \left( m + \frac{m^2}{2!} + \frac{m^3}{3!} + \ldots \right)
\]

\[
= me^m \left( 1 + \frac{m}{1!} + \frac{m^2}{2!} + \ldots \right) = me^m e^{-m} = m.
\]

Hence \( \mu_1' = \text{Mean} = m \).

Second moment

\[
\mu_2' = E(X^2) = \sum_{r=0}^{\infty} r^2 \cdot \frac{m^r e^{-m}}{r!} = \sum_{r=0}^{\infty} \left\{ r(r-1) + r \right\} \frac{m^r e^{-m}}{r!}
\]

\[
= \sum_{r=0}^{\infty} \frac{m^r e^{-m}}{(r-2)!} + \sum_{r=0}^{\infty} \frac{m^r e^{-m}}{(r-1)!}
\]

\[
= m^2 e^m e^{-m} + m = m^2 + m.
\]

Similarly third and fourth moments about origin are

\[
\mu_3' = E(X^3) = \sum_{r=0}^{\infty} r^3 \cdot \frac{m^r e^{-m}}{r!} = \sum_{r=0}^{\infty} \left\{ r(r-1)(r-2) + 3r(r-1) + r \right\} \frac{m^r e^{-m}}{r!}
\]

\[
= m^3 + m^2 + m \quad \text{and}
\]

\[
\mu_4' = E(X^4) = \sum_{r=0}^{\infty} r^4 \cdot \frac{m^r e^{-m}}{r!}
\]
Moments about mean: \( \mu_n \) (\( n = 1, 2, 3, 4, \ldots \))

\[ \mu_1 = 0 \text{ always,} \]
\[ \mu_2 = \mu_2' - (\mu_1')^2 = m^2 + m - m^2 = m \]

Or, variance \( \sigma^2 = m \), standard deviation (S.D.) = \( \sqrt{m} \) and mean = (S.D.)^2.

\[ \mu_3 = \mu_3' - 3 \mu_2' \mu_1' + 2 (\mu_1')^3 = m^3 + 3m^2 + m - 3 (m^2 + m)m + 2 m^3 \]
\[ = m^3 + 3m^2 + m - 3m^3 - 3m^2 + 2m^3 = m. \]

\[ \mu_4 = \mu_4' - 4 \mu_3' \mu_1' + 6 \mu_2' \mu_1'^2 - 3 \mu_1'^4 \]
\[ = m^4 + 6m^3 + 7m^2 + m - 4[m^3 + m^2 + m]m + 6 [m^2 + m^3 - 3 m^4 , \]
\[ = 3m^2 + m. \]

Karl Pearson’s Coefficients: \( \beta_1, \beta_2, \gamma_1, \gamma_2 \)

\[ \beta_1 = \frac{\mu_3}{\mu_2^3} = \frac{m^2}{m^3} = \frac{1}{m}, \quad \beta_2 = \frac{\mu_4}{\mu_2^3} = \frac{3m^2 + m}{m^3} = 3 + \frac{1}{m}, \quad \gamma_1 = \sqrt{\beta_1} = \frac{1}{\sqrt{m}}, \quad \gamma_2 = \beta_2 - 3 = \frac{1}{m}. \]

Note: (i) For Poisson’s distribution mean = variance \( m > 0 \) always, so the distribution is always positive skew. (ii) Both \( \gamma_1, \gamma_2 \) tends to 0 as \( m \) tends to infinity.

3.5.2. Recurrence Relation for the Moments of Binomial Distribution. For Poisson’s distribution with mean \( m \),

\[ \mu_{k+1} = m \mu_k + m \frac{d \mu_k}{dm}, \text{ where } \mu_k \text{ is the } k^{th} \text{ moment about mean.} \]

Proof: Since \( \mu_k = \sum_{r=0}^{\infty} (r - m)^k \cdot p(r) = \sum_{r=0}^{\infty} (r - m)^k \cdot \frac{m^r e^{-m}}{r!}, \)

On differentiating with respect to \( m \), we get
\[
\frac{d\mu_k}{dm} = -r \sum_{r=0}^{\infty} (r-m)^k \frac{m^r e^{-m}}{r!} + \sum_{r=0}^{\infty} (r-m)^k \left\{ \frac{r m^{r-1} e^{-m} - m^r e^{-m}}{r!} \right\}
\]

\[
= -r \mu_{k-1} + \sum_{r=1}^{\infty} (r-m)^k e^{-m} m^{r-1} \frac{r}{r!} (r-m)
\]

Therefore \( m \frac{d\mu_k}{dm} + m r \mu_{k-1} = \sum_{r=0}^{\infty} (r-m)^k \frac{e^{-m} m^r}{r!} = \mu_{k+1}.\)

If we put \( k = 1, 2, 3, \ldots \) we have

\[
\mu_2 = m \mu_0 + m \frac{d\mu_1}{dm} = m, \quad \mu_3 = 2 m \mu_1 + m \frac{d\mu_2}{dm} = m, \quad \mu_4 = 3 m \mu_2 + m \frac{d\mu_3}{dm} = 3 m^2 + m,
\]

etc.

3.5.3. Moment Generating Function of Poisson Distribution.

Let \( p(X = r) = \frac{m^r}{r!} e^{-m} ; \quad r = 0, 1, 2, \ldots, \infty \) be a Poisson distribution. Then m.g.f. about origin: \( M_0(t) = E(e^{t r}) = \sum_{r=0}^{\infty} e^{t r} \cdot e^{-m} \frac{m^r}{r!} = e^{-m} \sum_{r=0}^{\infty} \left( \frac{m e^t}{r!} \right) = e^{-m} \cdot e^{me^t} = e^m(e^t - 1) \)

m.g.f. about mean \( m: \)

\( M_m(t) = E(e^{t(r-m)}) = e^{-m} E(e^{t r}) = e^{-m} M_0(t) = e^{-m} \cdot e^{-m(1-e^t)} = e^{m(e^t-1)}. \)

Generation of moments of Poisson’s Distribution from m.g.f. : Since

\( M_0(t) = 1 + m(e^t - 1) + \frac{m^2}{2!}(e^t - 1)^2 + \frac{m^3}{3!}(e^t - 1)^3 + \frac{m^4}{4!}(e^t - 1)^4 + \ldots \)

\[
= 1 + m(t + \frac{t^2}{2!} + \frac{t^3}{3!} + \ldots) + \frac{m^2}{2} (t + \frac{t^2}{2!} + \ldots)^2 + \frac{m^3}{6} (t + \frac{t^2}{2!} + \ldots)^3 + \frac{m^4}{24} (t + \frac{t^2}{2!} + \ldots)^4 + \ldots
\]

Here \( \mu'_1 = \text{coefficient of } t = m, \quad \mu'_2 = \text{coefficient of } t^2/2! = m + m^2, \)

\( \mu'_3 = \text{coefficient of } t^3/3! = m + 3 m^2 + m^3, \)

\( \mu'_4 = \text{coefficient of } t^4/4! = m + (3m^2 + 4m^3) + 6m^3 + m^4. \)

Also, \( \mu = \mu'_1^2 - (\mu'_1)^2 = m + m^2 - m^2 = \text{variance} \) and S.D. \( \sigma = \sqrt{\mu}. \)

Similarly \( \mu_2, \mu_4, \ldots \) etc may be obtained. Also the moment about mean can be obtained by expanding \( M_m(t). \)
Alternative Method: \( \mu' = \left[ \frac{d}{dt} \{M_0(t)\} \right]_{t=0} = \left[ \frac{d}{dt} e^{m(t-1)} \right]_{t=0} \)

On putting \( r = 1, 2, 3, \ldots \), we get

\[
\mu'_1 = \left[ \frac{d}{dt} e^{m(t-1)} \right]_{t=0} = [e^{-m} m e' e'']_{t=0} = m.
\]

\[
\mu'_2 = \left[ \frac{d^2}{dt^2} e^{m(t-1)} \right]_{t=0} = [e^{-m} m e' e'' (me')^2 + e^{-m} m e' e']_{t=0}
\]

\[
= [e^{-m} m e' e' (m^2 e' + m)]_{t=0} = m^2 + m, \text{ etc}
\]

3.5.4. Mode of Poisson Distribution. (The most probable number of success).

The value of \( r \) which has a greater probability than any other value is the mode of the Poisson’s distribution. Let the probability of \( r \) successes is greater than or equal to that of \( r - 1 \) or \( r + 1 \) successes, i.e.,

\[
p(r - 1) \leq p(r) \geq p(r + 1).
\]

i.e.,

\[
\frac{m^{r-1}}{(r-1)!} e^{-m} \leq \frac{m^r}{r!} e^{-m} \geq \frac{m^{r+1}}{(r+1)!} e^{-m}.
\]

Simplifying we get, \( m \geq r \) and \( r + 1 \geq m \Rightarrow r \leq m \) and \( m - 1 \leq r \Rightarrow m - 1 \leq r \leq m \).

Case 1. If \( m \) is a positive integer, then there are two modes \( (m - 1) \) and \( m \).

Case 2. If \( m \) is not a positive integer, then there is one mode is the integral value between \( (m - 1) \) and \( m \).

3.5.5. Illustrative Examples.

Example 1. For Poisson’s distribution show that

(i) \( M \sigma \gamma_1 \gamma_2 = 1 \) and (ii) \( \sqrt{\beta_1(\beta_2 - 3)m} \sigma = 1 \).

Solution. Since we know that for a Poisson’s distribution,

\[
\text{Mean} = M = m, \quad \sigma = \sqrt{m}, \quad \gamma_1 = 1/\sqrt{m}, \quad \gamma_2 = 1/m, \quad \beta_1 = 1/m, \quad \beta_2 = 3 + (1/m)
\]

\( \Rightarrow M \sigma \gamma_1 \gamma_2 = m \sqrt{m} (1/\sqrt{m})(1/m) = 1 \) and \( \sqrt{\beta_1(\beta_2 - 3)m} \sigma = (1/\sqrt{m})(3 + (1/m) - 3)m \sqrt{m} = 1 \).

Example 2. Criticize the statement: For Poisson’s distribution mean is 7 and SD is 6.

Solution. Since for a Poisson’s distribution

\[
\sqrt{\text{mean}} = \text{SD} \Rightarrow \sqrt{7} = 6. \text{ So the statement is not correct.}
\]
Example 3. Find the probability that almost 5 defective fuses will be found in a box of 200 fuses if experience shows that 2 percent of such fuses are defective.

Solution: Here we use Poisson’s distribution, here \( m = np = 200 \times (2/200) = 4 \) and 

\[
e^{-4} = 1 - 4 + \frac{1}{2} \cdot 4^2 - \frac{1}{6} \cdot 4^3 + \frac{1}{24} \cdot 4^4 - ... = 0.0183.\text{ Hence}
\]

\[
P(r \leq 5) = \sum_{r=0}^{5} \frac{e^{-4} 4^r}{r!} = e^{-4} (1 + 4 + \frac{4^2}{2} + \frac{4^3}{6} + \frac{4^4}{24} + \frac{4^5}{120}) = 0.7845.
\]

Example 4. 6 coins are tossed 6400 times. Using Poisson distribution, find the approximate probability of getting six heads \( x \) times and 2 times.

Solution: Let the coins be unbiased so the probability of getting a head = the probability of getting a tail for each coin = \((1/2)^6 = 1/64 = p\) (say). Then \( m = np = 6400 \times (1/64) = 100 \).

Now the Poisson’s distribution \( p(X = r) = \frac{m^r}{r!} e^{-m} = \frac{100^r}{r!} e^{-100} \).

Therefore the required probability is \( p(X = 2) = \frac{100^2}{2!} e^{-100} = 5000 e^{-100} \).

Example 5. If \( P(X= 0) = P (X = 1) = a \) in a Poisson’s distribution, show that \( a = 1/e \).

Solution: The Poisson’s distribution is, \( p(X = r) = \frac{m^r}{r!} e^{-m} \), \( r = 0, 1, 2, \ldots, \infty \). Then

\[
P(X = 0) = P (X = 1) \Rightarrow e^{-m} = me^{-m} \Rightarrow m = 1. \text{ Therefore } P(X = 0) = e^1 = a.
\]

Example 6. In a Poisson’s distribution with unity mean, show that the mean deviation from mean is \( 2/e \) times the standard deviation. i.e., \( E|X - 1| = 2/e \).

Solution: The Poisson’s distribution is, \( p(X = r) = \frac{m^r}{r!} e^{-m} \), \( r = 0, 1, 2, \ldots, \infty \). Here \( m = 1 \), so \( SD = 1 \). Therefore \( p(X = r) = \frac{e^{-1}}{r!} \). Now mean deviation from mean

\[
= \sum_{r=0}^{\infty} |r - m| p(r) = \sum_{r=0}^{\infty} \frac{|r - 1|}{r!} = \frac{e^{-1}}{e} \sum_{r=0}^{\infty} \frac{|r - 1|}{r!} = \frac{1}{e} \left[ 1 + \frac{1}{2!} + \frac{2}{3!} + \frac{3}{4!} + \frac{4}{5!} + ... \right] = \frac{1}{e} \sum_{r=1}^{\infty} \frac{1}{r} = 2/e.
\]

Example 7. A Poisson’s distribution has a double mode at \( r = 3 \) and at \( r = 4 \), what is the probability that \( r \) will have one or the other of these values.
Solution: The Poisson’s distribution is, \( p(X = r) = \frac{m^re^{-m}}{r!} \), \( r = 0, 1, 2, \ldots, \infty \). Here two modes are given then the mean should be an integer and modes are \( m - 1 \) and \( m \). Therefore, \( m - 1 = 3 \), i.e., \( m = 4 \). Now

\[
p(X = r = 3) = \frac{4^3}{3!}e^{-4} \quad \text{and} \quad p(X = r = 4) = \frac{4^4}{4!}e^{-4} = \frac{4^3}{3!}e^{-4}.
\]
Hence the required probability is

\[
p(X = 3 \text{ or } 4) = p(X = 3) + p(X = 4) = \frac{4^3}{3!}e^{-4} + \frac{4^3}{3!}e^{-4} = \frac{64}{3}e^{-4}.
\]

Example 8. Is the sum of two independent Poisson’s variates a Poisson’s?

Solution: Let \( x_1 \) and \( x_2 \) be two independent Poisson’s variates with parameters \( m_1 \), \( m_2 \) respectively. Then the moment generating functions for \( x_1 \) and \( x_2 \) are

\[
M_{x_1}(t) = e^{m_1(e^t-1)} \quad \text{and} \quad M_{x_2}(t) = e^{m_2(e^t-1)}.
\]
Therefore the m.g.f for \( (x_1 + x_2) \) is

\[
M(x_1 + x_2)(t) = E(e^{(x_1+x_2)t}) = E(e^{x_1t} \cdot e^{x_2t}) = E(e^{x_1t}) \cdot E(e^{x_2t}) = M_{x_1}(t) \cdot M_{x_2}(t)
\]

\[
= e^{m_1(e^t-1)} \cdot e^{m_2(e^t-1)} = e^{(m_1+m_2)e^t-(m_1+m_2)}.
\]
This is the m.g.f. of a Poisson distribution with parameter \( m_1 + m_2 \). Hence \( x_1 + x_2 \) has a Poisson’s distribution with mean \( m_1 + m_2 \).

Example 9. In a certain factory turning razor blades, there is a small chance \( 1/500 \) for any blade to be defective. The blades are in packets of 10. Use Poisson’s distribution to calculate the approximate number of packets containing no defective, one defective and two defective blades in a consignment of 10,000 packets.

Solution: Since the probability distribution is \( p(X = r) = \frac{m^re^{-m}}{r!} \), \( r = 0, 1, 2, \ldots, \infty \). Then

\[
\text{expected frequency} = Np(X = r) = N \frac{m^re^{-m}}{r!}, \quad r = 0, 1, 2, 3, \ldots, \infty.
\]
Here \( N = 1000 \), \( m = np = 10 \times (1/500) = 0.02 \). So that \( e^{-0.02} = 0.9802 \). Then the respective expected frequencies for number of packets containing no defective, one defective and two defective blades are (i) \( Ne^{-m} = 10000 \times 0.9802 = 9802 \),

(ii) \( mNe^{-m} = 0.02 \times 10000 \times 0.9802 = 196 \),

(iii) \( (1/2)m^2 Ne^{-m} = (1/2) \times 0.004 \times 10000 \times 0.9802 = 2 \).

Example 10. Fit a Poisson’s distribution to the following and calculate theoretical frequencies:
<table>
<thead>
<tr>
<th>Deaths</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequencies</td>
<td>22</td>
<td>60</td>
<td>15</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Solution:

<table>
<thead>
<tr>
<th>x</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>22</td>
<td>60</td>
<td>15</td>
<td>2</td>
<td>1</td>
<td>200</td>
</tr>
<tr>
<td>fx</td>
<td>0</td>
<td>60</td>
<td>30</td>
<td>6</td>
<td>4</td>
<td>100</td>
</tr>
<tr>
<td>(fx^2)</td>
<td>0</td>
<td>60</td>
<td>60</td>
<td>18</td>
<td>16</td>
<td>154</td>
</tr>
</tbody>
</table>

Here mean = \(\frac{100}{200} = \frac{1}{2} = 0.5\) and SD = \(\sqrt{[(154/200) - (0.5)^2]} = 0.72\).

Therefore \(e^{-m} = e^{-0.5} = 0.61\) and the theoretical frequency of \(r\) deaths is given by

\[
\frac{m^r}{r!}e^{-m} = 200 \times \frac{(0.5)^r}{r!} \cdot 0.61.
\]

Therefore the Poisson’s distribution will be

<table>
<thead>
<tr>
<th>(r)</th>
<th>(P(X = r))</th>
<th>(\text{Expected frequency } N \cdot P(X = r))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>(e^{-m} = 0.61)</td>
<td>200 \times 0.61 = 122</td>
</tr>
<tr>
<td>1</td>
<td>(me^{-m} = 0.305)</td>
<td>200 \times 0.305 = 61</td>
</tr>
<tr>
<td>2</td>
<td>((m^2/2!)e^{-m} = 0.0762)</td>
<td>200 \times 0.0762 = 15</td>
</tr>
<tr>
<td>3</td>
<td>((m^3/3!)e^{-m} = 0.0127)</td>
<td>200 \times 0.0127 = 2</td>
</tr>
<tr>
<td>4</td>
<td>((m^4/4!)e^{-m} = 0.0016)</td>
<td>200 \times 0.0016 = 0</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>(1.0055 = 1, \text{ nearly})</td>
</tr>
</tbody>
</table>

Example 11. In a book of 300 pages, a proof reader finds no error in 200 pages, in 75 pages one error on each page, in 20 pages two errors on each page and in 5 pages 3 errors on each page. Use Poisson distribution to these data calculate theoretical frequency

\([e^{-0.43} = 0.6505]\)

Solution: The observed frequency distribution is

<table>
<thead>
<tr>
<th>No. of errors per page</th>
<th>x</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of pages f</td>
<td>200</td>
<td>75</td>
<td>20</td>
<td>5</td>
<td>300</td>
<td></td>
</tr>
<tr>
<td>(fx)</td>
<td>0</td>
<td>75</td>
<td>40</td>
<td>15</td>
<td>130</td>
<td></td>
</tr>
</tbody>
</table>
Then Mean $m = 130/300 = 0.43$ and $e^{-0.43} = 0.6505$, so the theoretical frequency of $r$ deaths is given by $N \frac{m^r}{r!} e^{-m} = 300 \times \frac{(0.43)^r}{r!} 0.6505$. Therefore the Poisson’s distribution will be according to the table shown below:

<table>
<thead>
<tr>
<th>$r$</th>
<th>$P(X = r)$</th>
<th>Expected frequency $N.P(X = r)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$e^{-m} = 0.6505$</td>
<td>$300 \times 0.6505 = 195.15$</td>
</tr>
<tr>
<td>1</td>
<td>$me^{-m} = 0.2797$</td>
<td>$300 \times 0.2797 = 83.91$</td>
</tr>
<tr>
<td>2</td>
<td>$(m^2/2!)e^{-m} = 0.0601$</td>
<td>$300 \times 0.0601 = 18.03$</td>
</tr>
<tr>
<td>3</td>
<td>$(m^3/3!)e^{-m} = 0.0086$</td>
<td>$300 \times 0.0086 = 2.58$</td>
</tr>
<tr>
<td>Total</td>
<td>$0.9989 = 1$, nearly</td>
<td>$299.67 = 300$, nearly</td>
</tr>
</tbody>
</table>

3.6. Normal Distribution:

(A) Normal distribution as a limiting case of binomial distribution when $p \neq q$ but $p \approx q$:

(General Case). Let $N(q + p)^n$ be the binomial distribution where $p \neq q$ but $p \approx q$ or $|p - q|$ is small. Now the ratio of the frequencies $f(r)$ and $f(r + 1)$ of $r$ and $r + 1$ successes respectively is $\frac{f(r + 1)}{f(r)} = \frac{n C_{r+1} p^{r+1} q^{n-r-1}}{n C_r p^r q^{n-r}} = \frac{n-r}{r+1} \frac{p}{q}$. Now The frequency of $r$ successes is greater than the frequency of $(r + 1)$ successes, i.e. $f(r) > f(r + 1)$.

If $\frac{f(r + 1)}{f(r)} = \frac{n-r}{r+1} \frac{p}{q} < 1$ or if $np - rp < rq + q$ or if $rq + rp > nq - q$. Similarly the frequency of $r$ successes is also greater than that of $(r - 1)$ successes if $r > np - q$ or if $(p + q)r > nq - q$ or if $r < np + q$.

Since $r = np$ is a possible value and we assume that $np$ is an integer. There is no loss of generality as $n$ is to become large ultimately. If $y_0$ be the maximum frequency, then
\[ y_0 = f(np) = N^n C_n p^n q^{n-np} = N \cdot \frac{n!}{(np)!((nq)!)} p^n q^{nq}. \]

Then the frequency of \( np + x \) successes is

\[ y_x = f(np + x) = N \frac{n!}{(np + x)!((nq - x)!)} p^{np+x} q^{nq-x}. \]

Therefore

\[
\frac{y_x}{y_0} = \frac{(np)!/(nq)!}{(np + x)!((nq - x)!)} p^x q^{-x}. \]

Now for large \( n \), by James Stirling formula \( n! = e^{-n} n^{n+1/2} \sqrt{2\pi} \), we have

\[
\log \frac{y_x}{y_0} = -(np + x + \frac{1}{2}) \log\left(1 + \frac{x}{np}\right) - (nq - x + \frac{1}{2}) \log\left(1 - \frac{x}{nq}\right)
\]

\[
= -(np + x + \frac{1}{2}) \left(\frac{x}{np} - \frac{x^2}{2np^2} + \frac{x^3}{3n^2p^3} - \ldots\right) - (nq - x + \frac{1}{2}) \left(\frac{x}{nq} - \frac{x^2}{2nq^2} + \frac{x^3}{3n^2q^3} - \ldots\right)
\]

\[
= x\left(-1 - \frac{1}{2np} + 1 + \frac{1}{2nq}\right) + x^2 \left(-\frac{1}{2np} - \frac{1}{4n^2p^2} + \frac{1}{2nq} - \frac{1}{4n^2q^2}\right) + \ldots
\]

\[
+ x^3 \left(-\frac{1}{3n^2p^2} + \frac{1}{6n^3q^3} - \frac{1}{2nq^2} - \frac{1}{3n^2q^2} - \frac{1}{6n^2q^2}\right) + \ldots
\]

\[
= x \left(\frac{p-q}{2npq} + x^2 \frac{p^2 + q^2}{4n^2 p^2 q^2} - \frac{x^2}{2npq}\right) + \ldots \text{ terms of higher order.}
\]

Neglecting terms containing \( 1/n^2 \), we get

\[
\log \frac{y_x}{y_0} = -\frac{q-p}{2npq} x - \frac{x^2}{2npq}.
\]

Since \( p < 1, q < 1 \), \( q - p \) is very small in comparison to \( n \) and therefore the first term can be neglected. Hence

\[
\log \frac{y_x}{y_0} = -\frac{x^2}{2npq}. \]

Therefore

\[
y_x = y_0 e^{-x^2/2npq} = y_0 e^{-x^2/2\sigma^2}.
\]

(B) Normal distribution as a limiting case of binomial distribution when \( p = q \):

(Particular Case). Let \( N(q + p)^n \) be the binomial distribution where \( p = q \). If \( p = q \) then \( p = q =1/2 \) and consequently the binomial distribution is symmetric. Now the frequency of \( r \) successes is \( f(r) = N \cdot C_r \cdot p^r q^{n-r} = N \cdot C_r (1/2)^n \). Without loss of generality, we assume that \( n \) is an even integer, say \( n = 2k \). Since \( n \to \infty \), the frequency \( f(r) = N^{2k} C_r (1/2)^{2k} \) Now the ratio of the frequencies \( f(r) \) and \( f(r + 1) \) of \( r \) and \( r + 1 \) successes respectively is
\[
\frac{f(r+1)}{f(r)} = \frac{2k}{r+1} C_{r+1} = \frac{2k-r}{r+1} C_r.
\]

Now the frequency of \( r \) successes is greater than the frequency of \((r + 1)\) successes, i.e. \( f(r) > f(r+1) \).

If \( \frac{f(r+1)}{f(r)} = \frac{2k-r}{r+1} < 1 \) or if \( 2k - r < r + 1 \) or if \( r > k - 1/2 \). Similarly the frequency of \( r \) successes is also greater than that of \((r - 1)\) successes if \( r < k + 1/2 \). Thus we observed that if \( k - \frac{1}{2} < r < k + \frac{1}{2} \), the frequency corresponding to \( r \) successes will be greatest. Clearly \( r = k \) is the value of success corresponding to which the frequency is maximum. Suppose \( y_0 \) be the maximum frequency, then

\[
y_0 = N^{2k} C_k (1/2)^{2k} = N \cdot \frac{(2k)!}{(k)!(k)!} (1/2)^{2k}.
\]

Then the frequency of \( k + x \) successes is

\[
y_x = \frac{(2k)!}{(k+x)!(k-x)!} (1/2)^{2k}.
\]

Therefore

\[
\frac{y_x}{y_0} = \frac{(k)!(k)!}{(k+x)!(k-x)!} \frac{k(k-1)(k-2)...(k-x+1)}{(k+x)(k+x-1)...(k+1)} = \frac{(1-\frac{1}{k})...(1-\frac{x-1}{k})}{(1+\frac{1}{k})...(1+\frac{x}{k})}.
\]

\[
\log \frac{y_x}{y_0} = \left[ \log(1-\frac{1}{k}) + \log(1-\frac{2}{k}) + ... + \log(1-\frac{x-1}{k}) \right] - \left[ \log(1+\frac{1}{k}) + \log(1+\frac{2}{k}) + ... + \log(1+\frac{x}{k}) \right]
\]

On expanding each logarithmic terms and neglecting the higher powers of \( x/k \), we get (as \( k \to \infty \))

\[
\log \frac{y_x}{y_0} = \frac{1}{k} \{1 + 2 + 3 + ... + (x-1)\} - \frac{1}{k} \{1 + 2 + 3 + ... + (x-1) + x\}.
\]

\[
= -\frac{2}{k} \{1 + 2 + 3 + ... + (x-1)\} - \frac{x}{k} = -\frac{2}{k} \frac{(x-1)(x+1)}{2} - \frac{x}{k} = -\frac{x}{k} (x-1) - \frac{x}{k} = -\frac{x^2}{k}.
\]

Implies that, \( y_x = y_0 e^{-x^2/k} \). But in the case of binomial distribution \( \sigma^2 = npq = n \times \frac{1}{2} \times \frac{1}{2} = n/4 = k/2 \). Hence \( y_x = y_0 e^{-x^2/2\sigma^2} = y_0 e^{-x^2/2\sigma^2} \).

(c) Determination of \( y_0 \). Suppose that \( y_0 \) be such that the total probability may be 1,

\[
\int_{-\infty}^{\infty} y_0 e^{-x^2/2\sigma^2} dx = 1 \Rightarrow 2y_0 \int_{0}^{\infty} e^{-x^2/2\sigma^2} dx = 1.
\]

Put \( x/\sigma\sqrt{2} = t, dx = \sqrt{2}\sigma dt \), we get

\[
2\sqrt{2}\sigma y_0 \int_{0}^{\infty} e^{-t^2} dt = 1 \Rightarrow 2\sqrt{2}\sigma y_0 (1/2)\sqrt{\pi} = 1 \Rightarrow y_0 = \frac{1}{\sqrt{2}\pi}. \]

Hence \( y = \frac{1}{\sqrt{2}\pi} e^{-x^2/2\sigma^2} \) is the
standard form of normal distribution. If the total frequency is N, the corresponding normal distribution is $$y = \frac{N}{\sqrt{2\pi}} e^{-x^2/2\sigma^2}.$$ 

If the origin is changed to the point \((m = np = \text{mean}, 0)\), where \(x - m\) is the excess of the mean over the value chosen as origin, then the corresponding normal distribution is 

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2}.$$ 

(D) Definition. If the probability density function of a continuous random variable \(X\) is defined by 

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2}, -\infty < x < \infty.$$ Then it is called Normal variable and its distribution is called Normal (or Gaussian) Distribution, here \(m\) and \(\sigma\) are called the parameter of the distribution. A curve is given by 

$$y = y_0 e^{-(x-m)^2/2\sigma^2}$$ is said to be a normal curve, where origin is taken at mean then 

$$y = y_0 e^{-x^2/2\sigma^2}.$$ 

(E) Properties of Normal distribution.

(1) The curve is symmetrical about the \(y\)-axis.

Since if we replace \(x\) by \(-x\) in the equation of normal curve 

$$y = y_0 e^{-(x-m)^2/2\sigma^2}$$ then the equation remains unchanged. Hence the normal curve is symmetric about \(x\)-axis.

(2) The mean, median and mode coincide at the origin.

Mean = \(m = \mu_1' = \int_{-\infty}^{\infty} y_0 e^{-x^2/2\sigma^2} xdx = 0\), since the integrand is an odd function. Now if 

$$y = y_0 e^{-x^2/2\sigma^2} \Rightarrow \frac{dy}{dx} = -\frac{y_0}{\sigma^2} xe^{-x^2/2\sigma^2} \text{ and } \frac{d^2y}{dx^2} = -\frac{y_0}{\sigma^2} \left(1 - \frac{x^2}{\sigma^2}\right) e^{-x^2/2\sigma^2}.$$ 

Then \(\frac{dy}{dx} = 0 \Rightarrow x = 0\) \text{ and } \(\frac{d^2y}{dx^2} < 0\). Hence at \(x = 0\) \(y\) is maximum, i.e., \(x = 0\) is the mode of normal distribution.

Also, If \(M_d\) is the median of the normal distribution then 

$$\int_{M_d/\sigma\sqrt{2}}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2} dx = \frac{1}{2} \Rightarrow 2 \frac{1}{\sigma\sqrt{2\pi}} \int_{M_d/\sigma\sqrt{2}}^{\infty} e^{-t^2} \sigma\sqrt{2} dt = \frac{1}{2},$$ where \(x/\sigma\sqrt{2} = t, dx = \sqrt{2}\sigma dt, \) 

$$\int_{M_d/\sigma\sqrt{2}}^{\infty} e^{-t^2} dt = \frac{1}{2} \sqrt{\pi} = \int_{0}^{\infty} e^{-t^2} dt \Rightarrow \frac{M_d}{\sigma\sqrt{2}} = 0 \Rightarrow M_d = 0.$$ Hence the mean, median and mode coincide at the origin.
(3) The points of inflexion of the normal curve are given by \( x = \pm \sigma \).

Since if \( y = y_0 e^{-x^2/2\sigma^2} \Rightarrow \frac{dy}{dx} = -\frac{y_0}{\sigma^2} xe^{-x^2/2\sigma^2} \), \( \frac{d^2 y}{dx^2} = -\frac{y_0}{\sigma^4} \left(1 - \frac{x^2}{\sigma^2}\right) e^{-x^2/2\sigma^2} \).

Then \( \frac{d^2 y}{dx^2} = 0 \Rightarrow 1 - \frac{x^2}{\sigma^2} = 0 \Rightarrow x = \pm \sigma \). Now \( \frac{d^3 y}{dx^3} = -\frac{xy_0}{\sigma^4} \left(3 - \frac{x^2}{\sigma^2}\right) e^{-x^2/2\sigma^2} \neq 0 \) at \( x = \pm \sigma \).


Let \( y = \frac{1}{\sigma \sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2} \). Then

\[
\text{Mean} = m = \mu' = \int_{-\infty}^{\infty} xf(x)dx = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} xe^{-(x-m)^2/2\sigma^2} dx
\]

\[
= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (m + 2\sigma)e^{-z^2/2}dz, \text{ where } z = (x - m)/\sigma.
\]

\[
= \frac{m}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2}dz + \sigma \int_{-\infty}^{\infty} e^{-z^2/2}dz = \frac{2m}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-z^2/2}dz,
\]

(since the integrand is even in first term and odd in second term)

\[
= \frac{m}{\sqrt{\pi}} \int_{0}^{\infty} e^{-t} \cdot t^{-1/2}dt, \text{ where } t = z^2/2, \ dt = z \ dz.
\]

\[
= \frac{m}{\sqrt{\pi}} \int_{0}^{\infty} e^{-t} \cdot t^{(1/2)-1}dt = m \frac{1/2}{\sqrt{\pi}} = m.
\]

Variance: \( \text{Var}(X) = E(X - m)^2 = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} (x - m)^2 e^{-(x-m)^2/2\sigma^2} dx
\]

\[
= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2}dz, \text{ where } z = (x - m)/\sigma.
\]

\[
= \frac{\sqrt{2}\sigma^2}{\sqrt{\pi}} \int_{0}^{\infty} z^2 e^{-z^2/2}dz, \quad \text{(since the integrand is even)}
\]

\[
= 2\sigma^2 \frac{1/2}{\sqrt{\pi}} \int_{0}^{\infty} e^{-t}dt, \text{ where } t = z^2/2, \ dt = z \ dz.
\]
\[ = 2\sigma^2 \frac{t^{(3/2)-1}}{\sqrt{\pi}} \int_0^\infty e^{-t} \, dt = \frac{2\sigma^3}{\sqrt{\pi}} = \sigma^2. \]

Hence \( \sigma^2 \) is the variance and \( \sigma \) is the SD of normal distribution.

With mean \( m \) and variance \( \sigma \) normal distribution is denoted by \( N(m, \sigma) \).

Median: If \( M_d \) is the median of the normal distribution then
\[
\int_{-\infty}^{M_d} \frac{1}{\sigma \sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2} \, dx = \frac{1}{2} \Rightarrow \int_{-\infty}^{(M_d-m)/\sigma} e^{-z^2/2} \, dz = \frac{1}{2}, \text{ where } z = \frac{x-m}{\sigma}. \]

On comparing, we get \( (M_d - m)/\sigma = 0 \), implies that \( M_d = m \).

Mode: Now if
\[
y = f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2} \Rightarrow \frac{dy}{dx} = \frac{1}{\sigma \sqrt{2\pi}} \left( -\frac{x-m}{\sigma^2} \right) e^{-(x-m)^2/2\sigma^2} = 0 \Rightarrow x = m. \]

Also
\[
\frac{d^2y}{dx^2} = \frac{1}{\sigma^2 \sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2} + \frac{1}{\sigma \sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2} \left( -\frac{x-m}{\sigma^2} \right)^2 < 0 \text{ at } x = m.
\]

Hence at \( x = m \), \( y \) is maximum, i.e., \( x = m \) is the mode of normal distribution and \( \max y = 1/\sigma \sqrt{2\pi} \).

Quartiles: The first quartile \( Q_1 \) is given by
\[
\int_{-\infty}^{Q_1} \frac{1}{\sigma \sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2} \, dx = \frac{1}{4}. \text{ On solving this equation with the help of Normal Probability Integral Table, we have } Q_1 = \mu - 0.6745\sigma. \]

The third quartile \( Q_3 \) is given by
\[
\int_{-\infty}^{Q_3} \frac{1}{\sigma \sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2} \, dx = \frac{3}{4}. \text{ On solving the equation with the help of 'Normal Probability Integral Table' we have } Q_3 = \mu + 0.6745\sigma.
\]

Moments about the Origin: Odd moments.
\[
\mu'_{2n+1} = E(X)^{2n+1} = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^\infty x^{2n+1} e^{-x^2/2\sigma^2} \, dx = 0, \text{ [Since the integrand is an odd function].}
\]

Hence the odd moments about origin are zero, i.e., \( \mu'_{1} = \mu'_{3} = \mu'_{5} = \ldots = \mu'_{2n+1} = 0 \).
Even moments. \( \mu'_{2n} = E(X)^{2n} = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} x^{2n} e^{-x^2/2\sigma^2} \, dx \)

\[
= \frac{2}{\sigma \sqrt{2\pi}} \int_{0}^{\infty} x^{2n} e^{-x^2/2\sigma^2} \, dx, \text{ [Since the integrand is an even function].}
\]

Put \( \frac{x^2}{2\sigma^2} = t, \, dx = \frac{\alpha \, dt}{\sqrt{2} \sqrt{t}} \), we get

\[
\mu'_{2n} = \frac{2}{\sigma \sqrt{2\pi}} \int_{0}^{\infty} t^{n} \sigma^{2n} e^{-t} \, \alpha \, dt = \frac{2^n \sigma^{2n}}{\sqrt{\pi}} \int_{0}^{\infty} t^{n+1-1} e^{-t} \, dt = \frac{2^n \sigma^{2n}}{\sqrt{\pi}} \left( n + \frac{1}{2} \right).
\]

Similarly, \( \mu'_{2n-2} = \frac{2^{n-1} \sigma^{2n-2}}{\sqrt{\pi}} \left( n-1 + \frac{1}{2} \right) \). Then

\[
\frac{\mu'_{2n-2}}{\mu'_{2n-2}} = 2\sigma^2 \left( \frac{n+1}{n-1} \right) = 2\sigma^2 \left( n-\frac{1}{2} \right) = \sigma^2 \left( 2n-1 \right) \Rightarrow \mu'_{2n} = \sigma^2 \left( 2n-1 \right) \mu'_{2n-2}.
\]

This is the recurrence formula for even moments. On putting \( n = n-1, n-2, n-3, \ldots, 3, 2, 1 \), we have

\[
\mu'_{2n-2} = \sigma^2 \left( 2n-3 \right) \mu'_{2n-4}, \quad \mu'_{2n-4} = \sigma^2 \left( 2n-5 \right) \mu'_{2n-6}, \ldots, \quad \mu'_{4} = 3\sigma^2 \mu'_{2}, \quad \mu_{2} = \sigma^2 \mu'_{0} = \sigma^2.
\]

Substituting the values one by one, we get

\[
\mu'_{2n} = (2n-1)(2n-3)...3.1 \sigma^{2n}.
\]

Moments about the mean: Odd moments.

\[
\mu_{2n+1} = E(X - m)^{2n+1} = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} (x-m)^{2n+1} e^{-(x-m)^2/2\sigma^2} \, dx
\]

\[
= \frac{\sigma^{2n+1}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^{2n+1} e^{-z^2/2} \, dz \quad \text{[where } z = (x-m)/\sigma] = 0.\text{[Since the integrand is an odd function].}
\]

Hence the odd moments about mean are zero. i.e., \( \mu_{1} = \mu_{3} = \mu_{5} = \ldots = \mu_{2n+1} = 0. \)

Even moments. \( \mu_{2n} = E(X - m)^{2n} = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} (x-m)^{2n} e^{-(x-m)^2/2\sigma^2} \, dx \)
\[
\frac{\sigma^{2n}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^{2n} e^{-z^2/2} \, dz \quad \text{[where \( z = (x - m) / \sigma \)}
\]
\[
= \frac{2\sigma^{2n}}{\sqrt{2\pi}} \int_{0}^{\infty} z^{2n} e^{-z^2/2} \, dz \quad \text{[Since the integrand is an even function]}
\]
\[
= \frac{2\sigma^{2n}}{\sqrt{2\pi}} \int_{0}^{\infty} (2t)^n e^{-t^2/2} \, dt , \quad \text{where \( t = z^2/2 \)} = \frac{2\sigma^{2n}}{\sqrt{2\pi}} \int_{0}^{\infty} \frac{2^{n-1}}{2^{n-1} t^{n-1}} e^{-t^2/2} \, dt
\]
\[
= \frac{2^n \sigma^{2n}}{\sqrt{\pi}} \int_{0}^{\infty} t^{n-1} e^{-t} \, dt = \frac{2^n \sigma^{2n}}{\sqrt{\pi}} \left( \frac{n+1}{2} \right) .
\]

Similarly, \( \mu_{2n-2} = \frac{2^{n-1} \sigma^{2n-2}}{\sqrt{\pi}} \int_{n-1}^{n+1} \frac{1}{2} . \) Then

\[
\frac{\mu_{2n}}{\mu_{2n-2}} = 2\sigma^2 \frac{\frac{n+1}{2}}{\frac{n-1}{2}} = 2\sigma^2 \left( n - \frac{1}{2} \right) = \sigma^2 (2n - 1) \Rightarrow \mu_{2n} = \sigma^2 (2n - 1) \mu_{2n-2} .
\]

This is the recurrence formula for even moments. On putting \( n = n-1, n-2, n-3, \ldots, 3, 2, 1 \), we have
\[
\mu_{2n-2} = \sigma^2 (2n - 3) \mu_{2n-4}, \quad \mu_{2n-4} = \sigma^2 (2n - 5) \mu_{2n-6}, \quad \ldots, \quad \mu_4 = 3\sigma^2 \mu_2, \quad \mu_2 = \sigma^2 \mu_0 = \sigma^2 .
\]

Substituting the values one by one, we get
\[
\mu_{2n} = (2n - 1)(2n - 3) \ldots 3.1 \sigma^{2n} . \quad \text{In Particular} \quad \mu_1 = 0, \mu_2 = \sigma^2, \mu_3 = 0, \mu_4 = 3\sigma^4 .
\]

Note: Since the mean of the normal distribution is zero so \( \mu_{2n} = \mu_{2n}^{'} \) and \( \mu_{2n+1} = \mu_{2n+1}^{'} \).

Karl Pearson’s Coefficients: \( \beta_1, \beta_2, \gamma_1, \gamma_2 \)
\[
\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0, \quad \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3\sigma^4}{(\sigma^2)^2} = 3, \quad \gamma_1 = \sqrt{\beta_1} = 0, \quad \gamma_2 = \beta_2 - 3 = 0 .
\]

Hence the normal distribution is symmetrical \text{i.e.} there is no skewness in normal distribution and it is mesokurtic \text{i.e.,} Kurtosis of a normal distribution vanish; it is neither platykurtic nor leptokurtic.

Mean Deviation about Mean.
\[ E(|X - m|) = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{|x-m|}{e^{-(x-m)^2/2\sigma^2}} \, dx = \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{z \, e^{-z^2/2}}{\sqrt{\pi}} \, dz, \quad \text{where } z = (x - m)/\sigma. \]

\[ = \frac{\sigma}{\sqrt{2\pi}} \left[ \int_{-\infty}^{0} (-z)e^{-z^2/2} \, dz + \int_{0}^{\infty} ze^{-z^2/2} \, dz \right] = \frac{2\sigma}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-z^2/2} \, dz, \]

\[ = \frac{2\sigma}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-t^2/2} \, dt \quad \text{(where } t = z^2/2, \, dt = z \, dz) = \sigma \sqrt{\frac{2}{\pi}} = 0.7979\sigma = \frac{4}{5} \text{ appr.} \]

Quartile Deviation: \[ \text{Q.D.} = \frac{Q_3 - Q_1}{2} = \frac{(\mu + 0.6745\sigma) - (\mu - 0.6745\sigma)}{2} = 0.6745\sigma = \frac{2}{3}\sigma, \]

which is also called the probable error.


Let \( y = \frac{1}{\sigma \sqrt{2\pi}} \, e^{-(x-m)^2/2\sigma^2} \).

Then m.g.f. About origin:

\[ M_0(t) = E(e^{ty}) = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} e^{ty} e^{-(x-m)^2/2\sigma^2} \, dx \]

\[ = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{t(y+\sigma)} e^{-z^2/2} \, dz, \quad \text{where } z = (x - m)/\sigma. \]

\[ = e^{tm} \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} \, dz + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z^2 - 2\zeta z\sigma + \zeta^2\sigma^2) + \frac{1}{2}(z^2\sigma^2)} \, dz \]

\[ = e^{\frac{mt}{2}} \frac{1}{\sqrt{\pi}} \int_{0}^{\infty} e^{-\frac{1}{2}(z - \zeta)^2} \, dz, \quad \text{(since the integrand is even)} \]

\[ = e^{\frac{mt + 1}{2}\sigma^2} \, \frac{1}{\sqrt{\pi}} \int_{0}^{\infty} e^{-y} \, \frac{1}{y^{1/2}} \, dy, \quad \text{where } (1/2)(z - t\sigma)^2 = y, \,(z - t\sigma)dz = dy. \]

\[ = e^{\frac{mt + 1}{2}\sigma^2} \, \frac{1}{\sqrt{\pi}} \int_{0}^{\infty} e^{-y} \, \frac{1}{y^{1/2}} \, dy = e^{\frac{mt + 1}{2}\sigma^2}. \]

Then m.g.f. About mean:
\[ M_n(t) = E(e^{t(x-m)}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{zt} e^{-(x-m)^2/2\sigma^2} dx \]

\[ = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{z^2/2} e^{-z^2/2} dz , \text{ where } z = (x - m)/\sigma. \]

\[ = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(z^2-2xz + x^2 + \frac{1}{2}z^2 - x^2)} dz \]

\[ = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-u^2} du , \text{ where } u = z - t\sigma = e^{t^2/2\sigma^2}. \]

Generation of moments by m.g.f.:

\[ M_n(t) = e^{t^2/2\sigma^2} = 1 + \frac{1}{2} t^2 \sigma^2 + \left( \frac{1}{3!} \right) \left( \frac{1}{2} t^2 \sigma^2 \right)^2 + \left( \frac{1}{3!} \right) \left( \frac{1}{2} t^2 \sigma^2 \right)^3 + ... \quad \text{Now} \]

\[ \mu_{2n+1} = \text{coefficients of } \left( \frac{t^{2n+1}}{(2n+1)!} \right) \text{ in } M_n(t) = 0 \quad \text{and} \]

\[ \mu_{2n} = \text{coefficients of } \left( \frac{t^{2n}}{(2n)!} \right) \text{ in } M_n(t) = \left( \frac{\sigma^{2n} (2n)!}{2^n n!} \right) = \frac{2n(2n-1)(2n-2)...3.1\sigma^{2n}}{n(n-1)(n-2)...3.2.1.2^n} \]

\[ = \frac{2^n n(n-1)(n-2)...3.2.1/(2n-1)(2n-3)...3.1\sigma^{2n}}{n(n-1)(n-2)...3.2.1.2^n} = (2n-1)(2n-3)...3.1\sigma^{2n} = \frac{(2n)!}{n!} \left( \frac{1}{2} \sigma^2 \right)^n. \]

3. Illustrative Examples.

Example 1. Two normal universe have the same total frequency but the standard deviation of one is k-time that of the other, show that the maximum frequency of the first is 1/k that of the other.

Solution: Let the total number of frequencies be N and \( \sigma \) and \( k\sigma \) be the standard deviations of the two universes. Hence the equations to the two normal curves are

\[ y = y_0 e^{-x^2/2\sigma^2} \quad \text{and} \quad y = y_0 e^{-x^2/2k^2\sigma^2}. \]

According the question if the total frequencies for both are the same, then

\[ y_0 \int_{-\infty}^{\infty} e^{-x^2/2\sigma^2} dx = y_0 \int_{-\infty}^{\infty} e^{-x^2/2k^2\sigma^2} dx \]

Or, \( y_0 \int_{0}^{\infty} e^{-x^2/2\sigma^2} dx = y_0 \int_{0}^{\infty} e^{-x^2/2k^2\sigma^2} dx \).
Put \( x/\sigma \sqrt{2} = u \) in left side and \( x/k \sigma \sqrt{2} = v \) in right side we have

\[
y_0 \int_0^\infty e^{-u^2/2\sigma^2} \, du = y_0 k \int_0^\infty e^{-v^2/2\sigma^2} \, dv \Rightarrow y_0 = ky_0'.
\]

But \( y_0 \) or \( y_0' \) are the frequencies corresponding to the mean or median or mode (as all the three averages coincide). Hence the maximum frequency of first is \( 1/k \) that of the other.

Example 2. In case of normal distribution, prove that \( \mu_{2n+2} = \sigma^2 \mu_{2n} + \sigma^3 \frac{d\mu_{2n}}{d\sigma} \).

Proof: Since \( \mu_{2n} = E(X - m)^{2n} = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} (x - m)^{2n} e^{-(x-m)^2/2\sigma^2} \, dx \)

On differentiating with respect to \( \sigma \), we get

\[
\frac{d\mu_{2n}}{d\sigma} = -\frac{1}{\sigma^2 \sqrt{2\pi}} \int_{-\infty}^{\infty} (x - m)^{2n} e^{-(x-m)^2/2\sigma^2} \, dx
\]

\[
+ \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} (x - m)^{2n} e^{-(x-m)^2/2\sigma^2} \left\{ -\frac{1}{2} (x-m)^2 \left( -\frac{2}{\sigma^3} \right) \right\} \, dx.
\]

\[
\Rightarrow \frac{d\mu_{2n}}{d\sigma} = -\frac{1}{\sigma^2 \sqrt{2\pi}} \int_{-\infty}^{\infty} (x - m)^{2n} e^{-(x-m)^2/2\sigma^2} \, dx + \frac{1}{\sigma^4 \sqrt{2\pi}} \int_{-\infty}^{\infty} (x - m)^{2n+2} e^{-(x-m)^2/2\sigma^2} \, dx.
\]

\[
\Rightarrow \frac{d\mu_{2n}}{d\sigma} = -\frac{1}{\sigma} \mu_{2n} + \frac{1}{\sigma^3} \mu_{2n+2} \quad \Rightarrow \quad \sigma^3 \frac{d\mu_{2n}}{d\sigma} = -\sigma^2 \mu_{2n} + \mu_{2n+2}.
\]

\[
\Rightarrow \mu_{2n+2} = \sigma^2 \mu_{2n} + \sigma^3 \frac{d\mu_{2n}}{d\sigma}.
\]

Example 3. Prove that for the normal distribution, the quartile deviation, the mean deviation from mean and the standard deviation are approximately in the ratio 10: 12: 15.

Solution: Since we know that for a normal distribution

\[
\text{S.D.} = \sigma, \quad \text{Q.D.} = \frac{2}{3} \sigma, \quad \text{M.D.} = \frac{4}{5} \sigma
\]

\[
\Rightarrow \text{Q.D.} : \text{M.D.} : \text{S.D.} = \frac{2}{3} \sigma : \frac{4}{5} \sigma : \sigma = 10\sigma : 12\sigma : 15\sigma = 10 : 12 : 15.
\]
Example 4. 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 variance of the normal 
distribution?

Solution: Since the mean and standard deviation of normal distribution are \( m \) and \( \sigma \) 
respectively. Given that first moment about 10 is given by 
\[
\mu'_1(10) = E(X - 10) = E(X) - 10 = m - 10 = 40 \Rightarrow m = 40 + 10 = 50
\]
Again, since mean = 50, then we have the fourth moment about 50 = fourth moment about 
mean = \( \mu'_4(50) = \mu_4 = 48 \) (given) = \( 3\sigma^4 \Rightarrow \sigma^2 = 4 \) = Variance.

Example 5. For a normal distribution with mean 2 and standard deviation 3, find the value 
of a variate such that the probability of the intervals from the mean to 
that value is 0.4114. Given that \( t = 1.35 \).

Solution. Let the random variable \( X \) has the p.d.f.
\[
y = f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2}. \text{Let the required value be } x, \text{ then}
\]
\[
P(2 < X < x) = \frac{1}{\sigma \sqrt{2\pi}} \int_{2}^{x} e^{-(x-m)^2/2\sigma^2} \, dx \quad \text{(Since given that mean } m = 2 \text{ and } \sigma = 3). \]

On putting \( t = \frac{x - m}{\sigma} = \frac{x - 2}{3} \), we get
\[
P(2 < X < x) = \frac{1}{\sqrt{2\pi}} \int_{0}^{t} e^{-t^2} \, dt = 0.4114 \text{ (given). Hence } x = 3t + 2 = 3 \cdot 1.35 + 2 = 6.05.
\]

Example 6. In an intelligence test administered to 1000 children, the average score is 42 and 
standard deviation is 24. Find (i) the number of children whose score exceeds 60 \( \text{ (ii) the number of children with score lying between 20 and 40.} \) It is given that if 
\[
f(t) = \frac{1}{\sqrt{2\pi}} \int_{0}^{t} e^{-t^2/2} \, dt, \text{ then } f(0.75) = 0.2734, f(0.91) = 0.3184, f(0.08) = 0.0319.
\]

Solution: Let, \( t = \frac{x - m}{\sigma} = \frac{60 - 42}{24} = 0.75 \), then \( P(t > 0.75) = 0.5 - 0.2734 = 0.2266, \)
\[
P(x > 60) = 0.2266. \text{ Therefore,}
\]
Required number of children = \( N \cdot P(X > 60) = 1000 \times 0.2266 = 226.6. \)
for $x = 20$, $t = \frac{x-m}{\sigma} = \frac{20-42}{24} = -0.91$ and for $x = 40$, $t = \frac{x-m}{\sigma} = \frac{40-42}{24} = 0.08$. Then

$$P[20 < X < 40] = P(-0.91 < t < -0.08) = P(0.08 < t < 0.91)$$

$$= P(0 < t < 0.91) - P(0 < t < 0.08) = 0.3184 - 0.0319 = 0.2865.$$

Therefore required no. of children $= 0.2865 \times 1000 = 286.5$.

Example 7. For a normal variate $X$, mean $= 12$ and standard deviation $= 2$. find $P(9.6 < x < 13.8)$. Given that for $x/\sigma = 0.9$, $A = 0.3159$ and for $x/\sigma = 1.2$, $A = 0.3849$.

Solution: It is given that $m = 12$ and $\sigma = 2$, therefore when $x = 9.6$

$$t = \frac{x-m}{\sigma} = \frac{9.6-12}{2} = -1.2$$

and when $x = 13.8$, $t = \frac{x-m}{\sigma} = \frac{13.8-12}{2} = 0.9$. Therefore the required probability

$$P(9.6 < X < 13.8) = \frac{1}{2\sqrt{2\pi}} \int_{9.6}^{13.8} e^{-\left(\frac{1}{2}(x-12)^2\right)/4} \, dx = \frac{1}{\sqrt{2\pi}} \int_{-1.2}^{0.9} e^{-\left(\frac{1}{2}t^2\right)} \, dt,$$

where $t = (x - 12)/2$

$$= \frac{1}{\sqrt{2\pi}} \int_{-1.2}^{0} e^{-\left(\frac{1}{2}t^2\right)} \, dt + \frac{1}{\sqrt{2\pi}} \int_{0}^{0.9} e^{-\left(\frac{1}{2}t^2\right)} \, dt,$$

$$= \frac{1}{\sqrt{2\pi}} \int_{0}^{1.2} e^{-\left(\frac{1}{2}t^2\right)} \, dt + \frac{1}{\sqrt{2\pi}} \int_{0}^{0.9} e^{-\left(\frac{1}{2}t^2\right)} \, dt,$$

$$= 0.3849 + 0.3159 = 0.7008.$$

Example 8. Assume the mean height of soldiers to be 68.22 inches with a variance 108(in)$^2$. How many soldiers in a regiment of 1000 would you expect to be over 6 feet tall? Given that the area under the standard normal curve between $t = 0$ and $t = 0.35$ is 0.1368 and between $t = 0$ and $t = 1.15$ is 0.3746.

Solution: It is given that $m = 68.22$ inches and $\sigma^2 = 10.8$ (in)$^2$, i.e., $\sigma = 3.28$ inches. Therefore

$$t = \frac{x-m}{\sigma} = \frac{72-68.22}{3.28} = 1.15 \text{ (where } x = 6' = 72'') \text{. Since it is given that the area under the standard normal curve between } t = 0 \text{ and } t = 1.15 \text{ is } 0.3746 \text{. So the area under the standard normal curve between } t = 1.5 \text{ and } t = \infty \text{ is } 0.5 - 0.3746 = 0.1254.$$

Hence number of soldiers who are over 6 feet tall $= 1000 \times 0.1254 = 125.4$.

Example 9. In a normal distribution 31% of the items are under 45 and 8% items are over 64. Find the mean and standard deviation. Given that if $f(t) = \frac{1}{\sqrt{2\pi}} \int_{0}^{t} e^{-x^2/2} \, dx$, then $f(0.5) = 0.19$, $f(1.4) = 0.42$. 

\[ f(t) = \frac{1}{\sqrt{2\pi}} \int_{0}^{t} e^{-x^2/2} \, dx \]
Solution: Let \( m \) and \( \sigma^2 \) are the mean and variance of normal distribution. It is given that \( P(X < 45) = 31\% = 0.31 \) and \( P(X < 64) = 8\% = 0.08 \).

Therefore \( P\left( \frac{X - m}{\sigma} \leq \frac{45 - m}{\sigma} \right) = 0.31 \) and \( P\left( \frac{X - m}{\sigma} \geq \frac{64 - m}{\sigma} \right) = 0.08 \) or \( P(t \leq \frac{45 - m}{\sigma}) = 0.31 \) and \( P\left( t \geq \frac{64 - m}{\sigma} \right) = 0.08 \), where \( t = \frac{X - m}{\sigma} \).

Since it is given that \( 0.19 = f(0.5) = \int_{-\infty}^{0.5} e^{-t^2/2} dt = \int_{-0.5}^{0} e^{-t^2/2} dt \),

\[
= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0.5} e^{-(1/2)t^2} dt + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-0.5} e^{-(1/2)t^2} dt . \text{ Therefore}
\]

\[
\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-0.5} e^{-(1/2)t^2} dt = 0.5 - 0.19 = 0.31, \text{ i.e., } P(t \leq -0.5) = 0.31. \text{ Again given that}
\]

\[
0.42 = f(1.4) = \int_{0}^{1.4} e^{t^2/2} dt = \int_{0}^{\infty} e^{t^2/2} dt - \int_{1.4}^{\infty} e^{t^2/2} dt . \text{ Therefore}
\]

\[
\frac{1}{\sqrt{2\pi}} \int_{1.4}^{\infty} e^{-(1/2)t^2} dt = 0.5 - 0.42 = 0.08, \text{ i.e. } P(t \geq 1.4) = 0.08. \text{ On comparing we have}
\]

\[
\frac{45 - m}{\sigma} = -0.5 \text{ and } \frac{64 - m}{\sigma} = 1.4, \text{ which implies that } m = 50 \text{ and } \sigma = 10.
\]

Example 10. Fit a normal curve from the following data:

<table>
<thead>
<tr>
<th>Length(cm)</th>
<th>8.60</th>
<th>8.59</th>
<th>8.58</th>
<th>8.57</th>
<th>8.56</th>
<th>8.55</th>
<th>8.54</th>
<th>8.53</th>
<th>8.52</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>9</td>
<td>10</td>
<td>8</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Solution:

<table>
<thead>
<tr>
<th>( x )</th>
<th>( f )</th>
<th>( \xi = x - A, A = 8.56 )</th>
<th>( \xi^2 )</th>
<th>( f\xi )</th>
<th>( f\xi^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.60</td>
<td>2</td>
<td>0.04</td>
<td>0.0016</td>
<td>0.08</td>
<td>0.0032</td>
</tr>
<tr>
<td>8.59</td>
<td>3</td>
<td>0.03</td>
<td>0.0009</td>
<td>0.09</td>
<td>0.0027</td>
</tr>
<tr>
<td>8.58</td>
<td>4</td>
<td>0.02</td>
<td>0.0004</td>
<td>0.08</td>
<td>0.0016</td>
</tr>
<tr>
<td>8.57</td>
<td>9</td>
<td>0.01</td>
<td>0.0001</td>
<td>0.09</td>
<td>0.0009</td>
</tr>
<tr>
<td>8.56</td>
<td>10</td>
<td>0.00</td>
<td>0.0000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>8.55</td>
<td>8</td>
<td>-0.01</td>
<td>0.0001</td>
<td>-0.08</td>
<td>0.0008</td>
</tr>
<tr>
<td>8.54</td>
<td>4</td>
<td>-0.02</td>
<td>0.0004</td>
<td>-0.08</td>
<td>0.0016</td>
</tr>
<tr>
<td>8.53</td>
<td>1</td>
<td>-0.03</td>
<td>0.0009</td>
<td>-0.03</td>
<td>0.0009</td>
</tr>
<tr>
<td>8.52</td>
<td>1</td>
<td>-0.04</td>
<td>0.0016</td>
<td>-0.04</td>
<td>0.0016</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>42</strong></td>
<td></td>
<td><strong>0.0060</strong></td>
<td><strong>0.11</strong></td>
<td><strong>0.0133</strong></td>
</tr>
</tbody>
</table>

Then mean $m = \frac{1}{42} \sum f \xi = 8.56 + \frac{0.11}{42} = 8.5626$ and standard deviation $\sigma = \sqrt{\frac{\sum f \xi^2}{\sum f} - \left( \frac{\sum f \xi}{\sum f} \right)^2} = \frac{0.0133}{42} - \left( \frac{0.11}{42} \right)^2 = 0.0175 \text{ cm (appr.)}$

Hence the required normal curve is: $y = \frac{N}{\sigma \sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}} = 9.8 e^{-0.163(x-8.563)^2}$.

3.7. Rectangular Distribution or Uniform distribution.

A probability distribution function in which the probability density function $f(x)$ is constant over the entire range of variable $X$, say $a \leq x \leq b$ and zero elsewhere is said to be rectangular distribution or uniform distribution.

This distribution is so called since the curve $y = f(x)$ describes a rectangle over the X-axis between the ordinates at $x = a$ and $x = b$. This implies that $X$ is a continuous variable. Let X be a random variable in the range $a$ to $b$ with p.d.f. $f(x)$ which is constant in $a \leq x \leq b$ and zero elsewhere. Then

$$\int_{-\infty}^{\infty} f(x) dx = 1 \Rightarrow \int_{a}^{b} f(x) dx = 1 \Rightarrow f(x) \int_{a}^{b} dx = 1 \Rightarrow f(x) = \frac{1}{b-a}$$

Definition. A rectangular distribution with parameter $a$ and $b$ $(a < b)$ if given by p.d.f.

$$f(x) = \frac{1}{b-a} , \quad a \leq x \leq b$$

$$= 0 , \quad \text{ elsewhere}$$

Distribution Function of Rectangular Distribution. The distribution function $F(x)$ of rectangular distribution is given by $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x) dx = \int_{a}^{x} \frac{1}{b-a} dx$. Hence
\[ F(x) = \begin{cases} \frac{x-a}{b-a}, & a \leq x < b \\ 1, & x \geq b \end{cases} \]

3.7.1. Constants of the Rectangular Distribution

Moments about origin:

\[
\mu_r = E(X^r) = \int_a^b x^r f(x) \, dx = \int_a^b x f(x) \, dx = \left[ \frac{x^2}{b-a} \right]^b_a = \frac{b^2 - a^2}{2(b-a)} \quad \text{Mean = } \frac{a+b}{2}. 
\]

\[
\mu_2 = E(X^2) = \int_a^b x^2 f(x) \, dx = \int_a^b x^2 \frac{1}{b-a} \, dx = \left[ \frac{x^3}{3(b-a)} \right]^b_a = \frac{b^3 - a^3}{3(b-a)} = \frac{a^2 + ab + b^2}{3}, \quad \text{and so on}
\]

\[
\mu_r = \left[ \frac{x^{r+1}}{(r+1)(b-a)} \right]^b_a = \frac{b^{r+1} - a^{r+1}}{(r+1)(b-a)} \quad \text{Moments about } a:
\]

\[
\mu_r'(a) = E(X-a)^r = \int_a^b (x-a)^r f(x) \, dx = \left[ \frac{(x-a)^{r+1}}{(r+1)(b-a)} \right]^b_a = \frac{(b-a)^{r+1}}{(r+1)(b-a)} = \frac{(b-a)^r}{r+1}.
\]

In particular, \( \mu_1'(a) = \frac{b-a}{2} \), \( \mu_2'(a) = \frac{(b-a)^2}{3} \), \( \mu_3'(a) = \frac{(b-a)^3}{4} \), \( \mu_4'(a) = \frac{(b-a)^4}{5} \).

Moments about mean: \( \mu_1 = 0 \) always,

\( \mu_2 = \text{variance} \)

\[
\mu_2' - (\mu_1')^2 = \frac{a^2 + ab + b^2}{3} - \left( \frac{b+a}{2} \right)^2 = \frac{4b^2 + 4ab^2 + 4a^2 - 3b^2 - 3a^2 - 6ab}{12} = \frac{(b-a)^2}{12}.
\]

Standard deviation, \( \sigma = \frac{b-a}{\sqrt{12}} \).

Also, \( \mu_2 = \mu_2'(a) - [\mu_1'(a)]^2 = \frac{(b-a)^2}{3} - \left( \frac{b-a}{2} \right)^2 = \frac{(b-a)^2}{12} \).

\[
\mu_3' = \mu_3'(a) - 3\mu_2'(a)\mu_1'(a) + 2[\mu_1'(a)]^3 = \frac{(b-a)^3}{4} - 3\frac{(b-a)^2}{3}\left( \frac{b-a}{2} \right) + 2\left( \frac{b-a}{2} \right)^3 = 0.
\]

\[
\mu_4 = \mu_4'(a) - 4\mu_3'(a)\mu_1'(a) + 6\mu_2'(a)[\mu_1'(a)]^2 - 3[\mu_1'(a)]^4 = \frac{(b-a)^4}{80}.
\]

Karl Pearson’s Coefficients: \( \beta_1, \beta_2, \gamma_1, \gamma_2 \)
\[ \beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0, \quad \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{(b-a)^4 / 80}{[(b-a)^2 / 12]^2} = 1.8, \quad \gamma_1 = \sqrt{\beta_1} = 0, \quad \gamma_2 = \beta_2 - 3 = -1.2. \]

Thus rectangular distribution is symmetrical and platykurtic.

Quartile: The first quartile \( Q_1 \) is given by
\[
\int_a^b \frac{x}{b-a} \, dx = \frac{1}{4} \Rightarrow \left[ \frac{x}{b-a} \right]_a^b = \frac{1}{4} \Rightarrow Q_1 = a + \frac{1}{4} (b-a). \quad \text{The second quartile (median) } Q_2 (M_d) \text{ is given by}
\]
\[
\int_a^b \frac{x}{b-a} \, dx = \frac{1}{2} \Rightarrow \left[ \frac{x}{b-a} \right]_a^b = \frac{1}{2} \Rightarrow Q_2 = a + \frac{1}{2} (b-a). \]

The third quartile \( Q_3 \) is given by
\[
\int_a^b \frac{x}{b-a} \, dx = \frac{3}{4} \Rightarrow \left[ \frac{x}{b-a} \right]_a^b = \frac{3}{4} \Rightarrow Q_3 = a + \frac{3}{4} (b-a).
\]

3.7.2. Moment Generating Function of Rectangular Distribution.

Let \( f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0, & \text{otherwise} \end{cases} \). Then m.g.f. About origin:

\[
M_0(t) = E(e^{tx}) = \int_a^b \frac{e^{tx}}{b-a} \, dx = \frac{1}{b-a} \left[ \frac{e^{tx}}{t} \right]_a^b = \frac{e^{bt} - e^{at}}{t(b-a)}
\]

\[
= \left( 1 + bt + \frac{b^2 t^2}{2!} + \ldots + \frac{b^r t^r}{r!} + \ldots \right) - \left( 1 + at + \frac{a^2 t^2}{2!} + \ldots + \frac{a^r t^r}{r!} + \ldots \right)
\]

\[
= \frac{(b-a)t + \frac{(b^2-a^2)t^2}{2!} + \ldots + \frac{(b^r-a^r)t^r}{r!} + \ldots}{t(b-a)}.
\]

Hence \( \mu_r' = \text{coefficient of } \frac{t^r}{r!} \frac{(b^r-a^r)}{t(b-a)} \), for \( r = 0, 1, 2, 3, \ldots \).

3.7.3. Illustrative Examples.

Example 1. Find interquartile range, quartile deviation and its coefficient for the rectangular distribution \( f(x) = \frac{1}{b-a} \), \( a \leq x \leq b \).

Solution. Since, we have \( Q_1 = a + \frac{1}{4} (b-a) \), \( Q_3 = a + \frac{3}{4} (b-a) \), therefore
Interquartile Range = $Q_3 - Q_1 = (1/2)(b - a)$, Q.D. = $\frac{Q_3 - Q_1}{2} = \frac{1}{4}(b - a)$ and coefficient of Q.D. = $\frac{(b - a)/2}{2a + (b - a)} = \frac{(b - a)}{2(b + a)}$.

Example 2. For the rectangular distribution, where $f(x) = \frac{1}{b - a}$, $a \leq x \leq b$.

Find the mean deviation with respect to mean.

Solution: We know that for given rectangular distribution mean = $(a + b)/2$, therefore

Mean deviation from mean = $\int_a^b \left| x - \frac{a + b}{2} \right| f(x)dx = \frac{1}{b - a} \int_a^b \left| x - \frac{a + b}{2} \right| dx$, $a \leq x \leq b$

$= \frac{1}{b - a} \int_{-(b-a)/2}^{(b-a)/2} |y| dy = \frac{b - a}{4}$, where $y = x - \frac{a + b}{2}$.

Example 3. For the rectangular distribution $f(x) = \frac{1}{2a}$, $-a \leq x \leq a$. Find the m.g.f., even and odd moments.

Solution. $M_0(t) = E(e^{tx}) = \int_{-a}^a e^{tx}dx = \frac{1}{2a} \left[ e^{tx} \right]_{-a}^a = \frac{e^{at} - e^{-at}}{2a} = \frac{1}{at} \sinh at$

$= \frac{1}{at} \left[ at + \frac{a^3 t^3}{3!} + \frac{a^5 t^5}{5!} + \frac{a^7 t^7}{7!} + ... \right] = 1 + \frac{a^2 t^2}{3!} + \frac{a^4 t^4}{5!} + \frac{a^6 t^6}{7!} + ...$

Clearly all odd order moments $\mu_{2n+1}' = 0$ and even order moments are given by

$\mu_{2n}' = \frac{a^{2n}}{(2n+1)}$. Since here mean = $\mu_1' = 0$, therefore $\mu_n = \mu_n'$.

3.8. Exponential Distribution.

The probability distribution having the probability density function $f(x)$, defined by

$$f(x) = \begin{cases} 0, & x \leq 0 \\ \alpha^e, & x > 0 \end{cases}$$

is called exponential distribution with parameter $\theta$. 
Here \( \int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{0} f(x)dx + \int_{0}^{\infty} f(x)dx = 0 + \int_{0}^{\infty} \partial e^{-\theta x} dx = \left[-e^{-\theta x}\right]_{0}^{\infty} = 1 \) and the distribution function of Exponential Distribution is given by

\[
F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx = \int_{0}^{x} \partial e^{-\theta x} dx = \begin{cases} 
0, & x < 0 \\
1-e^{-\theta x}, & x \geq 0.
\end{cases}
\]

3.8.1. Constants of the Exponential Distribution

Moments about origin:

Mean = \( \mu_1' = E(X) = \int_{0}^{\infty} x f(x)dx = \int_{0}^{\infty} x \partial e^{-\theta x} dx = \theta \sqrt{\frac{2}{\theta^2}} = \frac{1}{\theta} \).

\[
\mu_2' = E(X^2) = \int_{0}^{\infty} x^2 f(x)dx = \int_{0}^{\infty} x^2 \partial e^{-\theta x} dx = \theta^3 \sqrt{\frac{3}{\theta^2}} = \frac{2}{\theta^2}, \text{ and so on}
\]

\[
\mu_r' = E(X^r) = \int_{0}^{\infty} x^r f(x)dx = \int_{0}^{\infty} x^r \partial e^{-\theta x} dx = \theta^{r+1} \sqrt{\frac{r+1}{\theta^{r+1}}} = \frac{r+1}{\theta^r}.
\]

Moments about mean: \( \mu_1 = 0 \) always, \( \mu_2 = \text{variance} = \mu_2' - (\mu_1')^2 = \frac{2}{\theta^2} - \left(\frac{1}{\theta}\right)^2 = \frac{1}{\theta^2} \).

Standard deviation, \( \sigma = \frac{1}{\theta} \). \( \mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2[\mu_1']^3 = \frac{6}{\theta^3} - \frac{3}{\theta^2} \cdot \frac{1}{\theta} + 2\left(\frac{1}{\theta}\right)^3 = \frac{2}{\theta^3} \)

\[
\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'[\mu_1']^2 - 3[\mu_1']^4 = \frac{9}{\theta^4}.
\]

Karl Pearson’s Coefficients: \( \beta_1, \beta_2, \gamma_1, \gamma_2 \)

\[
\beta_1 = \frac{\mu_2}{\mu_3} \left(\frac{2}{\theta^2}\right)^2 = 4, \quad \beta_2 = \frac{\mu_4}{\mu_2} = \frac{9}{\theta^4} \cdot \frac{1}{\theta^4} = 9, \quad \gamma_1 = \sqrt{\beta_1} = 2, \quad \gamma_2 = \beta_2 - 3 = 6.
\]

Quartiles: The first quartile \( Q_1 \) is given by
\[ Q_2 \int_0^\infty \partial e^{-\theta x} \, dx = \frac{1}{4} \Rightarrow \left[ -e^{-\theta x} \right]^\infty_0 = \frac{1}{4} \Rightarrow Q_1 = -\frac{1}{\theta} (\log_e 3 - \log_e 4) \]. The second quartile (median) \( Q_2 \) is given by \( (M_d) \) is given by \( Q_2 \int_0^\infty \partial e^{-\theta x} \, dx = \frac{1}{4} \Rightarrow \left[ -e^{-\theta x} \right]^\infty_0 = \frac{1}{4} \Rightarrow Q_2 = \frac{1}{\theta} \log_e 2 \). The third quartile \( Q_3 \) is given by

\[ Q_3 \int_0^\infty \partial e^{-\theta x} \, dx = \frac{3}{4} \Rightarrow \left[ -e^{-\theta x} \right]^\infty_0 = \frac{3}{4} \Rightarrow Q_3 = \frac{1}{\theta} \log_e 4. \]

Interquartile Range = \( Q_3 - Q_1 = (1/\theta) \log_e 4 + (1/\theta)(\log_e 3 - \log_e 4) = (1/\theta) \log_e 3 \) and Q.D. = \( \frac{Q_3 - Q_1}{2} = (1/2\theta) \log_e 3 \). Also coefficient of quartile deviation is

\[ \text{Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{(1/\theta) \log_e 3}{(1/\theta) \log_e 4 - (1/\theta)(\log_e 3 - \log_e 4)} = \frac{\log_e 3}{\log_e 16 - \log_e 3}. \]

Mean deviation from mean = \( \int_0^\infty x - \frac{1}{\theta} \partial e^{-\theta x} \, dx = \int_0^\infty \partial x - 1 \partial e^{-\theta x} \, dx = \frac{1}{\theta} \int_0^\infty y - 1 \partial e^{-y} \, dy, \) \( (y = 0) \)

\[ = \frac{1}{\theta} \int_0^\infty y \partial e^{-y} \, dy + \frac{1}{\theta} \int_0^\infty y \partial e^{-y} \, dy = \frac{1}{\theta} (e^{-1} + e^{-1}) = \frac{2}{\theta}. \]

3.8.2. Moment Generating Function of Exponential Distribution.

Let \( f(x) = \partial e^{-\theta x}, x \geq 0 \). Then m.g.f. About origin:

\[ M_0(t) = \text{E}(e^{tx}) = \int_0^\infty e^{tx} \partial e^{-\theta x} \, dx = \theta^{-t} \left[ e^{-\theta x} \right]^\infty_0 = \theta^t = \left( 1 - \frac{t}{\theta} \right)^{-1} \]

\[ = 1 + \frac{t}{\theta} + \frac{t^2}{\theta^2} + \frac{t^3}{\theta^3} + \ldots. \]

Hence \( \mu_r' = \text{coefficient of } \frac{t^r}{r!} = \frac{r!}{\theta^r} \), for \( r = 0, 1, 2, 3, \ldots. \)

3.8.3. Illustrative Examples

Example 1. A random variable has an exponential distribution with probability density function given by \( f(x) = \begin{cases} 3e^{-3x}, & x > 0 \\ 0, & x \leq 0. \end{cases} \) What is the probability, that X is not less than 4? Find the mean and standard deviation. Show that C.V. = mean/S.D. = 1.
Solution: We have to find out that $P(X \geq 4) = \int_{4}^{\infty} 3e^{-3x} dx = \left[ -e^{-3x} \right]_{4}^{\infty} = e^{-12}$.

Mean $= E(X) = \int_{4}^{\infty} 3xe^{-3x} dx = \frac{2}{3\pi} = \frac{1}{3}$, Variance $= \int_{4}^{\infty} \left( x - \frac{1}{3} \right)^2 e^{-3x} dx = \left( \frac{1}{3} \right)^2 = \frac{3}{3\pi^2} = \frac{1}{9} = \frac{1}{9}$ and Standard Deviation $= 1/3$. Hence C.V. $= 1$.

Example 2. The income tax of a man is exponentially distributed with the probability density function given by

$$f(x) = \begin{cases} \frac{1}{3} e^{-\frac{1}{3}x}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

What is the probability that his income will exceed Rs. 17,000 assuming that the income tax is lived at the rate of 15% on the income above Rs. 15000?

Solution. If the income exceeds Rs. 17,000 then income tax will exceed by 15% of 17000 - 15000, i.e. exceeds by Rs. (15x2000)/100 = 300. Hence the required probability is $P(X > 300) = \int_{300}^{\infty} \frac{1}{3} e^{-\frac{1}{3}x} dx = \left[ -e^{-\frac{1}{3}x} \right]_{300}^{\infty} = e^{-100}$.

3.9. Summary of Unit

(i) Binomial Distribution: The probability distribution of random variable $x$ is called Binomial distribution if only take non-negative values and its distribution is given by,

$$P(r) = p(X = r) = \binom{n}{r} p^r q^{n-r}, r = 0,1,2,3,...,n$$

Constants of the Binomial Distribution:

Moments about origin:

$$\mu_1' = \text{Mean} = np, \quad \mu_2' = npq + n^2 p^2, \quad \mu_3' = n(n-1)(n-2)p^3 + 3n(n-1)p^2 + np$$

and

$$\mu_4' = n(n-1)(n-2)(n-3)p^4 + 6n(n-1)(n-2)p^3 + 7n(n-1)p^2 + np$$

Moments about mean: $\mu_1 = 0$ always, $\mu_2 = npq = \text{variance} \sigma^2 = npq$ and S.D. $= \sqrt{npq}$, $\mu_3 = npq(2q - 1) = npq(q - p)$, $\mu_4 = npq[1 + 3(n - 2)pq]$.

Karl Pearson’s Coefficients:

$$\beta_1 = \frac{(1-2p)^2}{npq}, \quad \beta_2 = 3 + \frac{(1-6pq)}{npq}, \quad \gamma_1 = \frac{1-2p}{\sqrt{npq}}, \quad \gamma_2 = \frac{(1-6pq)}{npq}.$$

Renovsky Formula: (Recurrence Relation)
\[ \mu_{k+1} = pq \left( nk \mu_{k-1} + \frac{d\mu_k}{dp} \right), \text{ where } \mu_k \text{ is the } k^{th} \text{ moment about mean.} \]

Moment Generating Function of Binomial Distribution:

Then m.g.f. about origin: \( M_0(t) = (q + pe^t)^n \)

m.g.f. about mean np : \( M_{np}(t) = (q e^{pt} + pe^{qt})^n. \)

(ii) Poisson’s Distribution: The probability distribution of random variable \( x \) is called Poisson’s distribution if \( x \) can assume non-negative integral values only and its distribution is given by, 
\[
P(r) = p(X = r) = \{ \frac{m^r}{r!}, r = 0, 1, 2, 3, \ldots \text{n} \}.
\]

Constants of the Poisson Distribution:

Moments about origin:

\[ \mu_1' = \text{mean} = m, \quad \mu_2' = m^2 + m, \quad \mu_3' = m^3 + m^2 + m, \quad \mu_4' = m^4 + 6m^3 + 7m^2 + m \]

Moments about mean: \( \mu_1 = 0 \) always, \( \mu_2 = m = \text{variance } \sigma^2, \text{ S.D.} = \sqrt{m}, \text{ mean } = (\text{S.D.})^2. \)

\[ \mu_3 = m, \quad \mu_4 = 3m^2 + m. \]

Karl Pearson’s Coefficients: \( \beta_1 = 1/m, \quad \beta_2 = 3 + 1/m, \quad \gamma_1 = \frac{1}{\sqrt{m}}, \quad \gamma_2 = \frac{1}{m}. \)

Recurrence Relation: For Poisson’s distribution with mean \( m, \)

\[ \mu_{k+1} = mr \mu_{r-1} + m \frac{d\mu_r}{dm}, \text{ where } \mu_k \text{ is the } k^{th} \text{ moment about mean.} \]

Moment Generating Function of Poisson Distribution.

Then m.g.f. about origin : \( M_0(t) = e^{m(e^t-1)} \) and m.g.f. about mean \( m : M_m(t) = e^{(me^t-me-m)}. \)

(iii) Normal Distribution: If the probability density function of a continuous random variable \( X \) is defined by 
\[
y = \frac{1}{\sigma \sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2}, -\infty < x < \infty. \]

Then it is called Normal variable and its distribution is called Normal (or Gaussian) Distribution, here \( m \) and \( \sigma \) are called the parameter of the distribution.

(1) Normal curve is symmetrical about the y-axis.

(2) The mean, median and mode coincide at the origin.
(3) The points of inflexion of the normal curve are given by \( x = \pm \sigma \).

Constants and recurrence relation of the Normal Distribution.

Mean = \( \mu' = m \), \( \mu_2' = \text{variance} = \sigma^2 \) and S.D. = \( \sigma \).

Median = \( M_d = m \), Mode = \( m \), Quartiles : \( Q_1 = \mu - 0.6745\sigma \), \( Q_3 = \mu + 0.6745\sigma \).

Moments about the Origin : \( \mu'_{2n+1} = 0 \), \( \mu'_{2n} = (2n-1)(2n-3)\ldots 3.1\sigma^{2n} \).

Moments about the mean : \( \mu_{2n+1} = 0 \), \( \mu_{2n} = (2n-1)(2n-3)\ldots 3.1\sigma^{2n} \). Also \( \mu_n = \mu'_n \). Karl Pearson’s Coefficients: \( \beta_1 = 0 \), \( \beta_2 = 3 \), \( \gamma_1 = 0 \), \( \gamma_2 = 0 \).

Mean Deviation about Mean = \( \frac{4}{5} \) appr. Quartile Deviation = \( \frac{2}{3} \sigma \).

Moment Generating Function of Normal Distribution. m.g.f. About origin:

\[ M_0(t) = e^{\frac{mt}{2} - \frac{t^2\sigma^2}{2}} \] and m.g.f. About mean: \( M_m(t) = e^{\frac{t^2\sigma^2}{2}} \).

(iii) Rectangular Distribution : A rectangular distribution with parameter \( a \) and \( b \) (\( a < b \)) if given by p.d.f.

\[ f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{elsewhere} \end{cases} \]

Constants of the Rectangular Distribution

Moments about origin : Mean = \( \mu_1' = (a + b)/2 \), \( \mu_2' = (a^2 + ab + b^2)/3 \) etc.

Moments about mean: \( \mu_1 = 0 \) always, \( \mu_2 = \text{variance} = (b - a)^2/12 \), S.D. = \( \frac{b-a}{\sqrt{12}} \).

\[ \mu_3 = 0, \mu_4 = \frac{(b-a)^4}{80} \]. Karl Pearson’s Coefficients: \( \beta_1 = 0 \), \( \beta_2 = 1.8 \), \( \gamma_1 = 0 \), \( \gamma_2 = -1.2 \).

Quartiles : \( Q_1 = a + \frac{1}{4}(b-a) \), \( Q_2 = a + \frac{1}{2}(b-a) \), \( Q_3 = a + \frac{3}{4}(b-a) \).

Moment Generating Function of Rectangular Distribution. m.g.f. About origin:

\[ M_0(t) = \frac{e^{bt} - e^{at}}{t(b-a)} \]
(iv) Exponential Distribution: The probability distribution having the probability density function \( f(x) = \begin{cases} 0, & x \leq 0 \\ \frac{1}{\theta} e^{-\frac{x}{\theta}}, & x > 0 \end{cases} \) is called exponential distribution with parameter \( \theta \).

Constants of the Exponential Distribution: Moments about origin:

\[
\text{Mean} = \mu_1 = 1/\theta, \quad \mu_2 = 2/\theta^2. \quad \text{Moments about mean: } \mu_1 = 0 \text{ always, } \mu_2 = \text{variance} = 1/\theta^2 \quad \text{and}
\]
\[
\text{S. D.} = \frac{1}{\theta}, \quad \mu_3 = 2/\theta^3, \quad \mu_4 = \frac{9}{\theta^4}.
\]

Karl Pearson’s Coefficients: \( \beta_1 = 4, \quad \beta_2 = 9, \quad \gamma_1 = 2, \quad \gamma_2 = 6 \).

Quartiles: \( Q_1 = -\frac{1}{\theta} (\log_e 3 - \log_e 4), \quad Q_2 = \frac{1}{\theta} \log_e 2, \quad Q_3 = \frac{1}{\theta} \log_e 4 \).

Interquartile Range = \( Q_3 - Q_1 = (1/\theta) \log_e 3 \) and \( Q.D. = \frac{Q_3 - Q_1}{2} = (1/2\theta) \log_e 3, \quad Q.D. \)
\[
= \frac{\log_e 3}{\log_e 16 - \log_e 3}. \quad \text{Mean deviation from mean } = 2/\theta e.
\]

Moment Generating Function of Exponential Distribution.

\[
m.g.f. \quad \text{About origin: } M_0(t) = \left(1 - \frac{t}{\theta}\right)^{-1}
\]

3.10. Assignment

Q.1. In case of binomial Distribution, write an expression for (i) the probability of at most \( r \) successes (ii) the probability of at least \( r \) successes.

Q.2. The mean and variance of a Binomial distribution are 4 and 4/3 respectively. Find (i) the probability of 2 successes, (ii) the probability of more than two successes, (iii) the probability of 3 or more than three successes.

Q.3. A perfect cubical die is thrown a large number of times in sets of 8. The occurrence of 5 or 6 is called a success. In what proportion of the sets you expect 3 successes.

Q.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 would you expect it to give no even numbers?
Q 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?

Q 6. Show that if two symmetrical Binomial distributions of degree n (the same number of observations) are so superposed that the first term of the one coincides with the $(r + 1)^{th}$ term of the other, the distribution formed by adding superposed terms is a symmetrical Binomial distribution of degree $(n + 1)$.

Q 7. In a Poisson distribution probability for $x = 0$ is 10%. Find the mean, given that $\log_{10} 10 = 2.3026$.

Q 8. A car-hire-firm has two cars, which it hires, out day by day. The number of demands for a car on each day is distributed as Poisson distribution with mean 1.5. Calculate the proportion of days on which neither car is used and the proportion of days on which some demand is refused. [$e^{-1.5} = 0.2231$].

Q 9. A telephone switch board handles 600 calls on the average during a rush hour. The board can make a maximum of 20 connections per minute. Use Poisson distribution to estimate the probability that the board will be over taxed during any given minute. [$e^{-1} = 0.00004539$]

Q 10. If $p(X = 2) = 9p(X = 4) + 90p(X = 6)$ in the Poisson distribution, then find $E(X)$.

Q 11. If $X$ is a normal variate with mean 8 and standard deviation 4, find (i)$P(X \leq 5)$ and (ii) $P(5 \leq X \leq 10)$.

Q 12. Fit a normal curve from the following data:

<table>
<thead>
<tr>
<th>X</th>
<th>100</th>
<th>95</th>
<th>90</th>
<th>85</th>
<th>80</th>
<th>75</th>
<th>70</th>
<th>65</th>
<th>60</th>
<th>55</th>
<th>50</th>
<th>45</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>7</td>
<td>12</td>
<td>10</td>
<td>9</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Q 13. Fit a normal curve from the following data:

<table>
<thead>
<tr>
<th>Interval</th>
<th>60-62</th>
<th>63-65</th>
<th>66-68</th>
<th>69-71</th>
<th>72-74</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>5</td>
<td>18</td>
<td>42</td>
<td>27</td>
<td>8</td>
</tr>
</tbody>
</table>

Q 14. For the rectangular distribution $f(x) = 1$, $1 \leq x \leq 2$. Find arithmetic mean, geometric mean, harmonic mean and standard deviation and verify that AM > GM > HM.

Q 15. If families are selected at random in a certain thickly populated area and their annual income in excess of Rs. 4000 is treated as a random variable having an exponential
distribution \( f(x) = \frac{1}{2000} e^{-\frac{x}{2000}}, x > 0 \). What is the probability that 3 out of 4 families selected in the area have income in excess of Rs. 5000?

3.11. Check your progress

Q 1. By differentiating the identity \( \sum_{r=0}^{n} C_r p^r q^{n-r} = (q + p)^n \); \( q = 1 - p \) with respect to \( p \) and then multiplying by \( p \), show that \( \mu' = np \).

Q 2. The number of males in each 106 eight pig litters was found and they are given by the following frequency distribution:

<table>
<thead>
<tr>
<th>Number of Male per litter</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>0</td>
<td>5</td>
<td>9</td>
<td>22</td>
<td>25</td>
<td>26</td>
<td>14</td>
<td>4</td>
<td>1</td>
<td>106</td>
</tr>
</tbody>
</table>

Assuming that the probability of an animal being male or female is even i.e., \( p = q = \frac{1}{2} \) and frequency distribution follows the Binomial law, calculate the expected frequencies.

Q 3. In litters of 4 mice the number of litters which contained 0, 1, 2, 3, 4 females were noted. The figures are given in the table below:

<table>
<thead>
<tr>
<th>Number of female mice</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of litters</td>
<td>8</td>
<td>32</td>
<td>34</td>
<td>24</td>
<td>5</td>
<td>103</td>
</tr>
</tbody>
</table>

If the chance of obtaining a female in a single trial is assumed constant, estimate this constant of unknown probability. Find also expected frequencies.

Q 4. Find Binomial distribution if \( \beta_1 = \frac{1}{36}, \beta_2 = \frac{35}{12} \).

Q 5. Compute mode of a Binomial distribution with \( p = \frac{1}{4} \) and \( n = 7 \).

Q 6. Show that a measure of skewness of the Binomial distribution is given by \( \frac{q - p}{(npq)^{1/2}} \)

and its kurtosis is \( 3 + \frac{1 - 6pq}{npq} \).
Q 7. Let $X$ be a Binomially distributed random variable with mean 10 and variance 5. Show that

(a) $P(X > 6) = (1/2)^{20} \sum_{r=7}^{20} C_r$ and (b) $P(3 < X < 12) = (1/2)^{20} \sum_{r=4}^{11} C_r$.

Q 8. The mean of a Binomial distribution is 20 and standard deviation is 4. Determine the distribution.

Q 9. If $x$ and $y$ are two independent Poisson’s variates where $P(X = 1) = P(X = 2)$ and $P(Y = 2) = P(Y = 3)$, find the variance of $(X - 2Y)$.

Q 10. Find the mean and standard deviation for the table of deaths of women over 85 year old recorded in a three year period.

<table>
<thead>
<tr>
<th>No. of death recorded in a day</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of days</td>
<td>364</td>
<td>376</td>
<td>218</td>
<td>89</td>
<td>33</td>
<td>13</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Q 11. In 1,000 extensive sets of trials for an event of small probability the frequencies $f$ of the number $x$ of successes are found to be:

<table>
<thead>
<tr>
<th>$x$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f$</td>
<td>305</td>
<td>365</td>
<td>210</td>
<td>80</td>
<td>28</td>
<td>9</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Assuming it to be a Poisson’s distribution calculate its mean, variance and expected frequencies.

Q 12. If $X$ is a Poisson variate and $P(X = 1) = P(X = 2)$, find $P(X = 4)$ and $P(X \leq 4)$.

Q 13. If a Poisson distribution has a double mode $x = 1$ and $x = 2$, find $P(X = 1)$.

Q 14. If a Poisson distribution has a double mode $x = 4$ and $x = 5$, find probability that $x$ will have either of these values.

Q 15. In a distribution exactly normal. 7% of the items are under 35 and 89% are under 63. What are the mean and s.d. of the distribution.

Q 16. For a normal distribution with mean 1 and standard deviation 3, find the probability that $3.43 \leq x \leq 6.19$.

Q 17. The quartiles of a normal distribution are 8 and 14 respectively. Show that the mean and standard deviations are respectively 11 and 4.4.

Q 18. In a sample of 1000 cases, the mean of certain test is 14 and standard deviation is 2.5. Assuming the normality of the distribution find (i) how many candidates score between 12
and 15 (ii) how many score below 8 and (iii) the probability that a candidate selected at random will score above 15?

Q 19. Show that for the rectangular distribution \( dF(x) = dx, 0 \leq x \leq 1 \) the m.g.f is \( (e^t - 1)/t \). Hence or otherwise show that \( \mu_k' = 1/(k + 1) \), \( \mu_2 = 1/12 \). Also prove that mean deviation about mean is \( 1/4 \).

Q 20. The sales tax of a shopkeeper has an exponential distribution with p.d.f.

\[
f(x) = \begin{cases} 
\frac{1}{4}e^{-\frac{x}{4}}, & x > 0 \\
0, & x \leq 0.
\end{cases}
\]

If sales tax is levied at the rate of 5%, what is the probability that his sales exceed Rs. 10,000?

3.12. Point of Discussion

At the end of the unit student discuss or seek clarification on some points, if so mention the points:

A: ----------------------------------------------------------------------------------------------------------------------------------

----------------------------------------------------------------------------------------------------------------------------------

----------------------------------------------------------------------------------------------------------------------------------

B: ----------------------------------------------------------------------------------------------------------------------------------

----------------------------------------------------------------------------------------------------------------------------------

----------------------------------------------------------------------------------------------------------------------------------

C: ----------------------------------------------------------------------------------------------------------------------------------
3.13. Suggested Study Materials


========

UNIT–4: CURVE FITTING, CORRELATION AND REGRESSION

Structure

4.1. Introduction

4.2. Objectives

4.3. Curve Fitting
4.3.1. Most Plausible

4.3.2. Value Method of Least Square

4.3.3. Change of Origin

4.3.4. Curve of type $y = ab^x$ and $y = ax^b$.

4.3.5. Illustrative Examples

4.4. Correlation and Regression

4.4.1. Correlation: Karl Pearson’s Coefficients of Correlation

4.4.2. Coefficients of Correlation for Bivariate frequency Distribution

4.4.3. Correlation of Ranks

4.4.4. Illustrative Examples

4.4.5. Regression: Line of Regression

4.4.6. Covariance

4.4.7. Illustrative Examples

4.5. Partial and Multiple Correlation

4.5.1. Distribution of Variables.

4.5.2. Properties of Residuals.

4.5.3. Multiple and Partial correlation Coefficients:

4.5.4. Illustrative Examples

4.6. Summary of Unit

4.7. Assignment
4.1. **Introduction:** After the study of frequency distribution, measure of central tendencies, mean deviation, standard deviation, moment, skewness, kurtosis, theory of probability, mathematical expectation and moment generating functions and theoretical distribution. In this unit we shall confine ourselves to the study of method of least square, curve fitting, concepts of correlation and regression, coefficient of correlation, rank correlation, multiple and partial correlation. The fitting of curve to a given data is very important both from the point of view of theoretical and practical statistics. In theoretical statistics, the study of correlation and regression can be regarded as fitting of linear curves to be given bivariate or multi-variate frequency distributions. In practical statistics it enable us to get a close functional relation between x and y. These relations expressed by a polynomial, exponential or logarithmic and be fitted by using the principle of least square. We know that if (i) if, number of equations are equal to number of unknowns then such equations has unique solution (ii) if number of equations are less than number of unknowns then such equations has infinitely many solutions, but (iii) if number of equations are more than number of unknowns then the solution of such equations is not possible. To solve such equations perhaps Gauss used the method of least square in 1975 but it was named and first published in 1805 by Legendre and so this method is also known by principle of Legendre.

4.2: **Objectives:** After the end of the unit the student will be able to understand/know the

1. **Method of least square**
2. **Concepts of curve fitting and their applications**
3. **Concepts of correlation between two variates and their applications**
4. **Rank correlation and applications**
5. **Concept of Regression and their applications**
6. Abel to solve the problems based on above concepts

4.3: Curve Fitting. Curve Fitting means an expression of the relationship between two variables by algebraic equations on the basis of observed data. It is considered very important both from the point of view of theoretical and practical Statistics. In theoretical Statistics the lines of regression can be regarded as fitting of linear curves to the given bivariate values. In practical Statistics we are required to find a functional relation between $x$ and $y$ where the dependent variable $y$ is expressed as a function of the independent variable $x$, which may involve integral powers.

The general problem of finding equations of approximating curves which fit given sets of data is called curve fitting. The simplest curve that can be fitted to a number of points is the straight line. But no straight line passes exactly through all the points although a great many lines may be drawn which nearly do so. Similarly we can find curves of second degree, third degree etc., which may give the best representation of the points.

4.3.1. Most Plausible Value. Suppose we have a number of independent linear equations in $n$ unknowns say $x, y, z, ...$

$$a_1 x + b_1 y + c_1 z + ... = A_1, a_2 x + b_2 y + c_2 z + ... = A_2, \ldots, a_m x + b_m y + c_m z + ... = A_m,$$

where $a_r, b_r, c_r, \ldots$ and $A_r$ are constants. For $m = n$, we can find a unique set of values of $x, y, z, ...$ to satisfy all these equations. However if $m > n$, i.e., the number of equations is greater than the number of unknowns, there may exist no such solution. In such cases we try to find out those values of $x, y, z, ...$ which will satisfy the given system of equations as nearly as possible. The principle of least square asserts that these values are those which make $U$ a minimum where

$$U = \sum_{r=1}^{m} \left( a_r x + b_r y + c_r z + \ldots - A_r \right)^2.$$  Applying the conditions of minimum i.e.,

$$\frac{\partial U}{\partial x} = \frac{\partial U}{\partial y} = \frac{\partial U}{\partial z} = \ldots = 0,$$

we will get $n$ equations called as normal equations. When these equations are solved simultaneously, they give the values of $x, y, z, ...$. These are called the best or most plausible values. On calculating the second order partial derivatives and substituting the values of $x, y, z, ...$ thus obtained, we will see that the expression will become positive. Hence $U$ is minimum.

4.3.2. Method of Least Square: Suppose that we have $m$ observations are $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_m, y_m)$ of two variables $x$ and $y$ and we are required to fit a curve of the type

$$y = a + bx + bx^2 + cx^3 + \ldots + kx^n \quad \ldots \quad \ldots \quad \ldots \quad (1)$$
from these values. Now we have to determine the constants \( a, b, c, \ldots k \), such that it represents the curve of best fit of that degree. If \( m = n \), we can in general find a unique set of values satisfying the given system of equations. But when \( m > n \), we get \( m \) equations by substituting the different values of \( x \) and \( y \) in equation (1) and we are required to find only \( n \) constants. Therefore no such solution may exist to satisfy all \( m \) equations. We therefore try to find those values of \( a, b, \ldots, k \) which may give the best fit i.e., which may satisfy all the equations as nearly as possible. The principle of least squares asserts suitable method in such cases.

Substituting \( x_1, x_2, \ldots, x_m \) for \( x \) in equation (1), we get

\[
Y_1 = a + bx_1 + bx_1^2 + cx_1^3 + \ldots + kx_1^n \\
Y_2 = a + bx_2 + bx_2^2 + cx_2^3 + \ldots + kx_2^n \\
\quad \quad \ldots \quad \quad \ldots \quad \quad \ldots \\
Y_m = a + bx_m + bx_m^2 + cx_m^3 + \ldots + kx_m^n
\]

Here the quantities \( Y_1, Y_2, \ldots, Y_m \) and \( y_1, y_2, \ldots, y_m \) are called the expected values and observed values of \( y \) corresponding to the values of \( x_1, x_2, \ldots, x_m \) of \( x \). The difference \( R_r = y_r - Y_r \) for different values of \( r \) are called residuals. A measure of the "goodness of fit" of the curve to the given data is provided by the quantity \( R_1^2 + R_2^2 + \ldots + R_m^2 \), if this is small, the fit is good, if it is large the fit is bad.

"Of all curves approximating a given set of points, the curve having the property that \( R_1^2 + R_2^2 + \ldots + R_m^2 \) is minimum is called best fitting curve" and a curve having this property is said to fit the data in the least square sense and is called a least square curve.

Define a quantity \( U \) such that

\[
U = R_1^2 + R_2^2 + \ldots + R_m^2 = \sum_{r=1}^{m} (y_r - Y_r)^2 = \sum_{r=1}^{m} (y_r - a - bx_r - cx_r^2 - \ldots)^2.
\]

Here \( U \) is said to be sum of squares of residues. The principle of least square asserts that the constant \( a, b, c, \ldots, k \) are chosen in such a way so that sum of squares of residues is minimum. The conditions that \( U \) is maximum or minimum of is

\[
\frac{\partial U}{\partial a} = \frac{\partial U}{\partial b} = \frac{\partial U}{\partial c} = \ldots = \frac{\partial U}{\partial k} = 0.
\]

Which implies that

\[
\sum y = ma + b \sum x + \ldots + k \sum x^n \\
\sum xy = a \sum x + b \sum x^2 + \ldots + k \sum x^{n+1} \\
\sum x^2 y = a \sum x^2 + b \sum x^3 + \ldots + k \sum x^{n+2}
\]
... ... ...

\[ \sum x^n y = a \sum x^n + b \sum x^{n+1} + \ldots + k \sum x^{2n}, \]

here we removed the subscripts. These equations are called the \textit{normal equations} and are \((n + 1)\) in number and on solving these equation we have the values of \((n + 1)\) unknowns \(a, b, c, \ldots, k\). On calculating the second order partial derivatives and on putting these values, they give a positive value of the function. Hence \(U\) is minimum.

Particular cases. When \(n = 1\) and \(n = 2\) in these cases the curve to be fitted is a straight line \(y = a + bx\) and second degree parabola respectively the corresponding normal equations are

\[ \sum y = ma + b \sum x, \quad \sum xy = a \sum x + b \sum x^2 \quad \text{and} \quad \sum y = ma + b \sum x + c \sum x^2, \]
\[ \sum xy = a \sum x + b \sum x^2 + c \sum x^3, \quad \sum x^2 y = a \sum x^2 + b \sum x^3 + c \sum x^4 \]

respectively.

The values of \(\sum y, \sum x, \sum xy, \sum x^2 y, \sum x^2 \ldots\) etc. are calculated by means of a table as is used in the calculation of standard deviation or coefficient of correlation. Then the values of \(a, b, \ldots\) can be obtained.

4.3.2. Change of Origin. When the values of \(x\) are of equal interval \(i.e.,\) \(x, x + h, x + 2h, \ldots\) and \(m\) is odd say \(2n + 1\), the normal equations can be simplified by taking the origin of \(x\) at the middle of the values and the interval \(h\) may be taken as unit of measurement. Thus if \(a\) be the middle value, then \(u = (x - a)/h\) takes the values \(-n, \ldots, -1, 0, 1, \ldots, n\) and we get

\[ \sum u = \sum u^3 = \sum u^5 = 0 \quad \text{and} \quad \text{if } m \text{ is even, say } 2n \text{ we take the origin of } x \text{ at the mean of the middle pair of values and } h/2 \text{ as the new unit. The values of } m \text{ then become } -(2n - 1), -(2n - 3), \ldots, -3, -1, 1, 3, \ldots, (2n - 3), (2n - 1) \text{ and we get } \sum u = \sum u^3 = \sum u^5 = 0. \]

4.3.4. Curve of type \(y = ab^x\) and \(y = ax^b\). To fit the curves of these types we follow the following method. Let the curve to be fitted be \(y = ab^x\) or \(\log y = \log a + x \log b\). The normal equations are

\[ \sum \log y = m \log a + \log b \sum x, \quad \sum x \log y = \log a \sum x + \log b \sum x^2. \]

These equations are solved for \(\log a\) and \(\log b\) and then \(a\) and \(b\) are found by taking antilog. Similarly if the curve to be fitted be \(y = ax^b\) or \(\log y = \log a + b \log x\). The normal equations are

\[ \sum \log y = m \log a + b \sum \log x, \quad \sum \log x \log y = \log a \sum \log x + b \sum (\log x)^2. \]

These equations are solved for \(\log a\) \textit{and} \(b\) \textit{and} then \(a\) is found by taking antilog.
4.3.5. Illustrative Examples.

Example 1. Find the most plausible values of x, y and z from the following equations:

\[ x - y + 2z = 3, \ 3x + 2y - 5z = 5, \ 4x + y + 4z = 21, \ -x + 3y + 3z = 14. \]

Solution. \[ U = (x - y + 2z - 3)^2 + (3x + 2y - 5z - 5)^2 + (4x + y + 4z - 21)^2 + (-x + 3y + 3z - 14)^2 \]
The most plausible values will be given by \[ \frac{\partial U}{\partial x} = \frac{\partial U}{\partial y} = \frac{\partial U}{\partial z} = 0. \] On differentiating \( U \) partially with respect to \( x \), we get

\[ \frac{\partial U}{\partial x} = 2(x - y + 2z - 3) + 6(3x + 2y - 5z - 5) + 8(4x + y + 4z - 21) - (-x + 3y + 3z - 14) = 0 \]

\[ \Rightarrow 27x + 6y = 88. \]

Similarly on differentiating \( U \) partially with respect to \( y \) and \( z \), we get

\[ 6x + 15y + z = 70 \text{ and } y + 54z = 107 \] respectively. On solving these equations the most plausible values of \( x, y \) and \( z \) are \( x = 2.47, \ y = 3.55 \) and \( z = 1.92 \).

Example 2. Fit a straight line to the following data regarding \( x \) as the independent variable:

<table>
<thead>
<tr>
<th>( x )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y )</td>
<td>1.0</td>
<td>1.8</td>
<td>3.3</td>
<td>4.5</td>
<td>6.3</td>
</tr>
</tbody>
</table>

Solution. Let the straight line to be fitted to the data be \( y = a + bx \), then the normal equations are

\[ \sum y = ma + b \sum x, \ \sum xy = a \sum x + b \sum x^2. \] Now

<table>
<thead>
<tr>
<th>( x )</th>
<th>( y )</th>
<th>( xy )</th>
<th>( x^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1.8</td>
<td>1.8</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3.3</td>
<td>6.6</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>4.5</td>
<td>13.5</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>6.3</td>
<td>25.2</td>
<td>16</td>
</tr>
</tbody>
</table>
Here \( m = 5 \), \( \sum y = 16.9, \sum x = 10, \sum xy = 47.1, \sum x^2 = 30 \). Then by normal equations, we have \( 16.9 = 5a + 10b, 47.1 = 10a + 30b \). On solving these equations, we get \( a = 0.72, b = 1.33 \). Hence the required line is \( y = 0.72 + 1.33x \).

Example 3. Fit a second degree parabola to the following data:

<table>
<thead>
<tr>
<th>( x )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y )</td>
<td>1</td>
<td>5</td>
<td>10</td>
<td>22</td>
<td>38</td>
</tr>
</tbody>
</table>

Solution. Let the second degree parabola to be fitted to the data be \( y = a + bx + cx^2 \), then the normal equations are

\[
\sum y = ma + b \sum x + c \sum x^2 ,
\]

\[
\sum xy = a \sum x + b \sum x^2 + c \sum x^3
\]

and

\[
\sum x^2 y = a \sum x^2 + b \sum x^3 + c \sum x^4 .
\]

Now

<table>
<thead>
<tr>
<th>( X )</th>
<th>( Y )</th>
<th>( x^2 )</th>
<th>( x^3 )</th>
<th>( x^4 )</th>
<th>( xy )</th>
<th>( x^2y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>9</td>
<td>27</td>
<td>81</td>
<td>66</td>
<td>198</td>
</tr>
<tr>
<td>4</td>
<td>38</td>
<td>16</td>
<td>64</td>
<td>256</td>
<td>152</td>
<td>608</td>
</tr>
<tr>
<td>10</td>
<td>76</td>
<td>30</td>
<td>100</td>
<td>354</td>
<td>243</td>
<td>851</td>
</tr>
</tbody>
</table>

Here

\( m = 5, \sum y = 76, \sum x = 10, \sum x^2 = 30, \sum x^3 = 100, \sum x^4 = 354, \sum xy = 243, \sum x^2 y = 851 \). Then by normal equations, we have

\( 76 = 5a + 10b + 30c, 243 = 10a + 30b + 100c, 851 = 30a + 100b + 354c \). On solving these equations, we get \( a = 1.43, b = 0.24, c = 2.21 \). Hence the required parabola is

\( y = 1.43 + 0.24x + 2.21x^2 \).

Example 4. Fit a straight line to the following data regarding \( x \) as the independent variable:

<table>
<thead>
<tr>
<th>( x )</th>
<th>0</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
</tr>
</thead>
</table>
Solution. Here \( m = 6 \) even and the values of \( x \) are equally spaced \( h = 5 \). Therefore we take \( u = (x - 12.5)/2.5 \) and \( v = y - 20 \). Let the straight line to be fitted to the data be \( v = a + bu \), then the normal equations are

\[
\sum v = ma + b\sum u, \quad \sum uv = a\sum u + b\sum u^2 \quad \text{Now}
\]

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>u</th>
<th>v</th>
<th>uv</th>
<th>( u^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>12</td>
<td>-5</td>
<td>-8</td>
<td>40</td>
<td>25</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>-3</td>
<td>-5</td>
<td>15</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>17</td>
<td>-1</td>
<td>-3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>22</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>24</td>
<td>3</td>
<td>4</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>25</td>
<td>30</td>
<td>5</td>
<td>10</td>
<td>50</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>0</td>
<td>0</td>
<td>122</td>
<td>70</td>
</tr>
</tbody>
</table>

Now by normal equations, we have \( 0 = 6a, \ 122 = 70b \) implies that \( a = 0, \ b = 1.74 \). Hence the required line is \( v = 1.74u \) or \( y - 20 = 1.74[(x - 12.5)/2.5] \) or \( y = 0.7x + 11.28 \).

Example 5. Fit a second degree parabola to the following:

<table>
<thead>
<tr>
<th>x</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>1</td>
<td>1.8</td>
<td>1.3</td>
<td>2.5</td>
<td>6.3</td>
</tr>
</tbody>
</table>

Solution. Since \( m \) is odd and the values of \( x \) are equi-distant we take the origin for the \( x \) series at the middle value 2. Now let us put \( X = x - 2 \) and \( Y = y \), so that the curve of fit is \( Y = a + bX + cX^2 \).

<table>
<thead>
<tr>
<th>x</th>
<th>Y</th>
<th>X</th>
<th>Y</th>
<th>XY</th>
<th>( X^2 )</th>
<th>( X^3 )</th>
<th>( X^4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>-2</td>
<td>1</td>
<td>-2</td>
<td>4</td>
<td>4</td>
<td>-8</td>
</tr>
<tr>
<td>1</td>
<td>1.8</td>
<td>-1</td>
<td>1.8</td>
<td>-1.8</td>
<td>1</td>
<td>1.8</td>
<td>-1</td>
</tr>
<tr>
<td>2</td>
<td>1.3</td>
<td>0</td>
<td>1.3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2.5</td>
<td>1</td>
<td>2.5</td>
<td>2.5</td>
<td>1</td>
<td>2.5</td>
<td>1</td>
</tr>
</tbody>
</table>
Now the normal equations are \[ \sum Y = ma + b \sum X + c \sum X^2, \]
\[ \sum XY = a \sum X + b \sum X^2 + c \sum X^3, \]
\[ \sum X^2Y = a \sum X^2 + b \sum X^3 + c \sum X^4. \]

On substituting the values in these equations, we get

\[
12.9 = 5a + 10c, \quad 11.3 = 10a + 10b, \quad 33.5 = 10a + 34c.
\]

On solving these equations, we get \( a = 1.48, b = 1.13, c = 0.55 \). Hence the required parabola is \( Y = 1.48 + 1.13X + 0.55X^2 \) or \( y = 1.48 + 1.13(x - 2) + 0.55(x - 2)^2 \), or \( y = 1.48 - 1.07x + 0.55x^2 \).

Example 6. Fit a curve \( y = ab^x \) to the following data:

<table>
<thead>
<tr>
<th>( x )</th>
<th>( 2 )</th>
<th>( 3 )</th>
<th>( 4 )</th>
<th>( 5 )</th>
<th>( 6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y )</td>
<td>144</td>
<td>172.8</td>
<td>207.4</td>
<td>248.8</td>
<td>298.5</td>
</tr>
</tbody>
</table>

Solution. We have to fit the curve \( y = ab^x \), taking log both sides we have

\[
\log y = \log a + x \log b.
\]

So the normal equations are

\[
\sum \log y = m \log a + \log b \sum x, \quad \sum x \log y = \log a \sum x + \log b \sum x^2.
\]

Then

<table>
<thead>
<tr>
<th>( X )</th>
<th>( y )</th>
<th>( x^2 )</th>
<th>( \text{Logy} )</th>
<th>( x \log y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>144</td>
<td>4</td>
<td>2.1584</td>
<td>4.3168</td>
</tr>
<tr>
<td>3</td>
<td>172.8</td>
<td>9</td>
<td>2.2375</td>
<td>6.7125</td>
</tr>
<tr>
<td>4</td>
<td>207.4</td>
<td>16</td>
<td>2.3168</td>
<td>9.2672</td>
</tr>
<tr>
<td>5</td>
<td>248.8</td>
<td>25</td>
<td>2.3959</td>
<td>11.9795</td>
</tr>
<tr>
<td>6</td>
<td>298.5</td>
<td>36</td>
<td>2.4749</td>
<td>14.8494</td>
</tr>
<tr>
<td>20</td>
<td>90</td>
<td>115835</td>
<td>47.1254</td>
<td></td>
</tr>
</tbody>
</table>

Then by the normal equations we get

\[
11.5835 = 5 \log a + 20 \log b, \quad 47.1254 = 20 \log a + 90 \log b.
\]
Solving these equations and taking anti-logarithms, we get \( a = 100, \ b = 1.2 \) approximately. Hence the equation of the required curve is \( y = 100(1.2)^x \).

4.4. Correlation and Regression: In earlier units we have been mainly concerned with univariate universes only. In this section we shall discuss bivariate universes and in particular study the simultaneous variation of two variables for examples: height and weight of students in a class, ages of husband and wives, rainfall and crops etc. Let us consider the heights of the husbands and the wives at the time of marriage. If the height of the bridegroom is represented by \( x \) in general and that of bride by \( y \), then to each marriage there corresponds a pair of values \((x_1, y_1)\) of the variables \( x \) and \( y \). Now our object is to discover whether there is any connection between stature of husband \((x)\) and stature of wife \((y)\). Do tall men tend on the average to wed tall women, or do we find tall men choosing short women for wives just about as often as they choose tall women? Then we try to find out a relation between \( x \) and \( y \). Whenever two variables \( x \) and \( y \) are so related that a change in one is accompanied by change in the other in such a way that an increase in the one is accompanied by an increase or decrease in the other, then variables are said to be correlated. When the relation relationship is of quantitative nature, the approximate statistical tool for discovering and measuring the relationship and expressing it in brief formula is known as correlation.

Let the points \((x_r, y_r)\), \( r = 1, 2, 3, \ldots \) are the values of two variables \( x \) and \( y \) be plotted on a \( xy \)-plane and be represented by points \( P_1, P_2, P_3, \ldots \) . The values in the brackets are being taken the corresponding values of two variables. One of the variable is taken along the \( x \)-axis and the other along \( y \)-axis. Such a graphical representation is called a Scatter or Dot-diagram.

A universe every number of which bears one of the values of each of two variates is said to be bivariate. If the pair \((x_1, y_1)\) occurs \( f_1 \) times then \( f_1 \) is called the frequency of that pair. If the values are grouped according to class-intervals, we have a bivariate frequency distribution.

When the number of measurements becomes large, some class-interval is adopted for the measurement of such variables. A table can be obtained from the scatter-diagram by sub-dividing the coordinate area into equal rectangular compartments and then writing within each compartment is called the correlation table. For example if we take intervals 25-30, 30-35, 35-40, 40-45, 45-50 for \( x \) and the same for \( y \), then the correlation table is given by:
When an increase (or decrease) in one variate corresponds to an increase (or decrease) in the other, the correlation is said to be positive. It is negative when increase in one corresponds to decrease in the other or vice versa.

4.4.1. Correlation: Karl Pearson’s Coefficients of Correlation: The coefficients of correlation \( r \) between two variables \( X \) and \( Y \) is defined by the relation

\[
r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} = \frac{\sum xy}{n} \cdot \frac{p}{\sigma_x \sigma_y}.
\]

Where \( x \) and \( y \) are the deviations measured from their respective means and \( \sigma_x, \sigma_y \) being the standard deviations of these series. This is also called the product moment correlation coefficient. In many cases it is easier to work out after taking assumed means and in that case the formula of correlation coefficient becomes a different one. Let \( x \) and \( y \) be the deviations measured from the true means \( M_x \) and \( M_y \) of the two series and \( \xi \) and \( \eta \) be the deviations measured from the assumed means \( A_x \) and \( A_y \) respectively. Then \( \xi = X - A_x = X - M_x + M_x - A_x = x + d_x \), where \( d_x = M_x - A_x = \frac{\sum \xi}{n} \), similarly \( \eta = Y - A_y = Y - M_y + M_y - A_y = y + d_y \), where \( d_y = M_y - A_y = \frac{\sum \eta}{n} \). Then

\[
\sum \xi \eta = \sum (x + d_x)(y + d_y) = \sum (xy + xd_y + yd_x + d_x d_y)
\]

\[
= \sum xy + d_x \sum x + d_y \sum y + \sum d_x d_y
\]

\[
= \sum xy + \sum d_x d_y, \text{ since } \Sigma x = 0 = \Sigma y.
\]

Hence \( \sum xy = \sum \xi \eta - nd_x d_y = \sum \xi \eta - \frac{\sum \xi \sum \eta}{n} \).
Also, \( \sigma_x = \frac{\sum \xi^2 - \left( \frac{\sum \xi}{n} \right)^2}{n} \) and \( \sigma_y = \frac{\sum \eta^2 - \left( \frac{\sum \eta}{n} \right)^2}{n} \). Then the formula of coefficient is reduced to \( r = \frac{\sum \xi \eta - \frac{\sum \xi \sum \eta}{n}}{\sqrt{\left\{ \frac{\sum \xi^2 - \left( \frac{\sum \xi}{n} \right)^2}{n} \right\} \cdot \left\{ \frac{\sum \eta^2 - \left( \frac{\sum \eta}{n} \right)^2}{n} \right\}} \). This is known the short cut method of correlation coefficient. If we suppose that \( u = \frac{\xi}{h} \) and \( v = \frac{\eta}{h'} \), where \( h \) and \( h' \) are the scales of \( x \) and \( y \) series. Then

\[
 r = \left[ \frac{\sum \xi \eta - \frac{\sum \xi \sum \eta}{n}}{\sqrt{\left\{ \frac{\sum \xi^2 - \left( \frac{\sum \xi}{n} \right)^2}{n} \right\} \cdot \left\{ \frac{\sum \eta^2 - \left( \frac{\sum \eta}{n} \right)^2}{n} \right\}} \right] = \left[ \frac{\sum u v - \frac{\sum u \sum v}{n}}{\sqrt{\left\{ \frac{\sum u^2 - \left( \frac{\sum u}{n} \right)^2}{n} \right\} \cdot \left\{ \frac{\sum v^2 - \left( \frac{\sum v}{n} \right)^2}{n} \right\}}} \right] \]

is the step-deviation method of correlation coefficient.

4.4.2. Coefficients of Correlation for Bivariate frequency Distribution: For a bivariate frequency distribution the coefficient of correlation is given by

\[
 r = \frac{\sum f u v - \frac{\sum f u \sum f v}{\sum f}}{\sqrt{\left\{ \frac{\sum f u^2 - \left( \frac{\sum f u}{\sum f} \right)^2}{\sum f} \right\} \cdot \left\{ \frac{\sum f v^2 - \left( \frac{\sum f v}{\sum f} \right)^2}{\sum f} \right\}}} , \text{ where } f \text{ is the frequency in the frequency table.}
\]

Degree of correlation. When we calculate correlation coefficient \( r \), the following table shows the results of correlation.

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Result of correlation</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Perfect</td>
<td>+1</td>
<td>-1</td>
</tr>
<tr>
<td>2</td>
<td>High degree</td>
<td>+0.75</td>
<td>-0.75</td>
</tr>
<tr>
<td>3</td>
<td>Mediate Degree</td>
<td>b/w +0.5 to +0.75</td>
<td>b/w -0.5 to -0.75</td>
</tr>
<tr>
<td>4</td>
<td>Low Degree</td>
<td>b/w 0 to +0.5</td>
<td>b/w 0 to -0.5</td>
</tr>
<tr>
<td>5</td>
<td>No correlation</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Standard error and Probable error: If $r$ is the correlation coefficient then

$$S.D. = \frac{1-r^2}{\sqrt{n}} \quad \text{and} \quad P.E. = 0.6745 \frac{1-r^2}{\sqrt{n}}.$$ The P.E. error is used for testing the reliability of a particular value of $r$. The formula $r \pm P.E.$ gives two limits within which the coefficient of correlation always lies. If $r < P.E.$, there is a correlation and if $r < 6P.E.$ the correlation is significant.

Correlation coefficient lies between $-1$ to $+1$: If $x$ and $y$ denotes the deviations of the variates X and Y from their respective means, then by Schwarz’s Inequality

$$[E(xy)]^2 \leq E(x^2)E(y^2).$$ On dividing both sides by $E(x^2)E(y^2)$, we get $r^2 \leq 1$ or $-1 \leq r \leq +1.$ (see section 4.4.6)

4.4.3. Correlation of Ranks: An easier method than that of Karl Pearson’s method of calculation of correlation coefficient is given by Charls Spearman, which known as Spearman’s Rank Difference Method. In this method a group of $n$ individuals are arranged in order of merit in the possession of a certain characteristic. The same group would in general give different orders for different characteristics. In this method the only ranks are considered and so we call it the rank correlation coefficient in the characteristics for those group individuals. Assuming that no two individuals X and Y are equal in either classification, each of the individual takes the values 1, 2, 3, ..., $n$ and hence their arithmetic means are equal and each being $\overline{X} = \frac{1+2+3+...+n}{n} = \frac{n+1}{2} = \overline{Y}$. Let $d = X - Y = (X - \overline{X}) - (Y - \overline{Y}) = x - y$, where $x$ and $y$ are deviation from the mean. Therefore

$$\sum x^2 = \sum (X - \overline{X})^2 = \sum X^2 + \sum \overline{X}^2 - 2\sum X \overline{X} = \frac{n(n+1)(2n+1)}{6} + n\left(\frac{n+1}{2}\right)^2 - (n+1)\frac{n(n+1)}{2}$$

$$= \frac{n(n^2-1)}{12} = \sum y^2.$$

Now

$$\sum d^2 = \sum (x - y)^2 = \sum x^2 + \sum y^2 - 2\sum xy \Rightarrow \sum xy = \frac{1}{2}\left(\sum x^2 + \sum y^2 - \sum d^2\right) = \frac{1}{2}\left(\frac{n(n^2-1)}{6} - \sum d^2\right)$$

Hence

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} = \frac{1}{12} \frac{n(n^2-1) - \frac{1}{2} \sum d^2}{\frac{n(n^2-1)}{12} - \frac{6}{n(n^2-1)}} = 1 - \frac{6\sum d^2}{n(n^2-1)}.$$ In this formula is called Spearman’s coefficient of Rank Correlation.
Rank Correlation Coefficient lies between -1 and +1 including both the values. Since we know that \( r \) is a positive number and \( r \) is maximum if \( \sum d^2 \) is minimum. Now \( \sum d^2 \) is minimum if each \( d \) is zero i.e., so that \( \sum d^2 = 0 \). Hence maximum value of \( r \) is 1. Now \( r \) is minimum if \( \sum d^2 \) is maximum i.e., each of \( d \) is maximum, \( d \) will be maximum if the ranks of the \( n \) individuals are in opposite party as shown below:

\[
\begin{array}{ccccccc}
  x & : & 1 & 2 & 3 & \ldots & n-1 & n \\
  y & : & n & n-1 & n-1 & \ldots & 2 & 1
\end{array}
\]

Now two cases arise: When \( n \) is odd and when \( n \) is even.

Case I. Let \( n \) be odd and equal to \( 2r + 1 \). Then different \( d \)'s are

\[
(2r + 1 - 1), (2r + 1 - 1 - 2), \ldots, 4, 2, 0, -2, -4, \ldots, -(2r - 2), -2r
\]

or

\[
2r, (2r - 2), \ldots, 2, 0, -2, \ldots, -(2r - 2), -2r.
\]

Therefore

\[
\sum d^2 = 2[(2r)^2 + (2r - 2)^2 + 4^2 + 2^2] = 8[r^2 + (r - 1)^2 + 2^2 + 1^2] = (8/6)[r(r + 1)(2r + 1)].
\]

\[
\Rightarrow r = 1 - \frac{6\sum d^2}{n(n^2 - 1)} = 1 - \frac{8r(r + 1)(2r + 1)}{(2r + 1)(4r^2 + 4r + 1 - 1)} = 1 - 2 = -1.
\]

Case II. When \( n \) is even say \( 2r \). Then the values of \( d \) are

\[
(2r - 1), (2r - 3), \ldots, 3, 1, -1, -3, \ldots, -(2r - 3), -(2r - 1).
\]

Therefore

\[
\sum d^2 = 2[(2r - 1)^2 + (2r - 3)^2 + 3^2 + 1^2] = 2[(2r)^2 + (2r - 1)^2 + (2r - 2)^2 + 3^2 + 2^2 + 1^2 - (2r)^2 + (2r - 2)^2 + \ldots + 4^2 + 2^2] = (2/3)r(4r^2 - 1).
\]

\[
\Rightarrow r = 1 - \frac{6\sum d^2}{n(n^2 - 1)} = 1 - \frac{4r(4r^2 - 1)}{2r(4r^2 - 1)} = 1 - 2 = -1.
\]

Thus we get \(-1 \leq r \leq 1\), which are the required limits.

4.4.4. Illustrative Examples.

Example 1. The students got the following percentage of marks in Economics and Statistics:

\[
\begin{array}{ccccccc}
  x & : & 1 & 2 & 3 & \ldots & n-1 & n \\
  y & : & n & n-1 & n-1 & \ldots & 2 & 1
\end{array}
\]
Roll Nos.                  :  1  2  3  4  5  6  7  8  9  10
Marks in Economics :  78  36  98  24  75  82  90  62  65  40
Marks in Statistics :  84  51  91  60  68  62  86  58  53  47

Calculate the coefficient of correlation.

Solution. Let the marks of two subjects be denoted by X and Y respectively. Then the mean for X-series = M_x = 650/10 = 65 and the mean for Y-series = M_y = 660/10 = 66. If x and y are the deviations of X’s and Y’s from their respective means, then the data may be arranged in the following form:

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>x</th>
<th>y</th>
<th>x^2</th>
<th>y^2</th>
<th>xy</th>
</tr>
</thead>
<tbody>
<tr>
<td>78</td>
<td>84</td>
<td>13</td>
<td>18</td>
<td>159</td>
<td>324</td>
<td>234</td>
</tr>
<tr>
<td>36</td>
<td>51</td>
<td>-29</td>
<td>-15</td>
<td>841</td>
<td>225</td>
<td>435</td>
</tr>
<tr>
<td>98</td>
<td>91</td>
<td>33</td>
<td>25</td>
<td>1089</td>
<td>625</td>
<td>825</td>
</tr>
<tr>
<td>24</td>
<td>60</td>
<td>-40</td>
<td>-6</td>
<td>1600</td>
<td>36</td>
<td>240</td>
</tr>
<tr>
<td>75</td>
<td>68</td>
<td>10</td>
<td>2</td>
<td>100</td>
<td>4</td>
<td>20</td>
</tr>
<tr>
<td>82</td>
<td>62</td>
<td>17</td>
<td>-4</td>
<td>289</td>
<td>16</td>
<td>-68</td>
</tr>
<tr>
<td>90</td>
<td>86</td>
<td>25</td>
<td>20</td>
<td>625</td>
<td>400</td>
<td>500</td>
</tr>
<tr>
<td>62</td>
<td>58</td>
<td>-3</td>
<td>-8</td>
<td>9</td>
<td>64</td>
<td>24</td>
</tr>
<tr>
<td>65</td>
<td>53</td>
<td>0</td>
<td>-13</td>
<td>0</td>
<td>169</td>
<td>0</td>
</tr>
<tr>
<td>40</td>
<td>47</td>
<td>-26</td>
<td>-19</td>
<td>676</td>
<td>361</td>
<td>494</td>
</tr>
<tr>
<td>650</td>
<td>660</td>
<td>0</td>
<td>0</td>
<td>5398</td>
<td>2224</td>
<td>2704</td>
</tr>
</tbody>
</table>

Therefore \( r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} = \frac{2704}{\sqrt{5398 \times 2224}} = 0.78 \text{appr.} \)

Example 2. A computer while calculating the correlation coefficient between two variates x and y from 25 pairs of observations obtained the following constants:
n = 25, \( \Sigma x = 125, \) \( \Sigma x^2 = 650, \) \( \Sigma y = 100, \) \( \Sigma y^2 = 960, \) \( \Sigma xy = 508. \) It was however, later discovered at the time of checking that he had copied down two pairs as

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>While the correct values were</th>
<th>x</th>
<th>y</th>
<th>Obtained the correct value of the correlation coefficient.</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>14</td>
<td></td>
<td>8</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td></td>
<td>6</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

Solution. On account of the mistake, there will be no change in \( \Sigma x, \) \( \Sigma y \) and \( \Sigma x^2. \) But there will be change in \( \Sigma y^2 \) and \( \Sigma xy. \) Instead of old value in \( \Sigma y^2, \) \( 14^2 + 6^2 = 232 \) the new value \( 12^2 + 8^2 = 208 \) is to be substituted. Hence correct value of \( \Sigma y^2 = 960 - 232 + 208 = 936. \) Also instead of old value in \( \Sigma xy, \) \( 6 \times 14 + 8 \times 6 = 132, \) the new value \( 8 \times 12 + 6 \times 8 = 144 \) is to be substituted. Hence the correct value of \( \Sigma xy = 508 - 132 + 144 = 520. \)

\[
\text{Hence } r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} = \frac{520}{\sqrt{650 \times 936}} = 0.666.
\]

Example 3. Find the coefficient of correlation between the values of \( X \) and \( Y: \) (short cut method)

<table>
<thead>
<tr>
<th>X:</th>
<th>1</th>
<th>3</th>
<th>5</th>
<th>7</th>
<th>8</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y:</td>
<td>8</td>
<td>12</td>
<td>15</td>
<td>17</td>
<td>18</td>
<td>20</td>
</tr>
</tbody>
</table>

Solution. Let the assumed mean for \( X \) be 7 and for \( Y \) be 15. Then

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>( \xi )</th>
<th>( \eta )</th>
<th>( \xi \eta )</th>
<th>( \xi^2 )</th>
<th>( \eta^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>-6</td>
<td>-7</td>
<td>42</td>
<td>36</td>
<td>49</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>-4</td>
<td>-3</td>
<td>12</td>
<td>16</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>-2</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>17</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>18</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>3</td>
<td>5</td>
<td>15</td>
<td>9</td>
<td>25</td>
</tr>
</tbody>
</table>

Total  -8  0  72  66  96
Example 4. **Find the coefficient of correlation for the following table: (step deviation method)**

<table>
<thead>
<tr>
<th>x</th>
<th>10</th>
<th>14</th>
<th>81</th>
<th>22</th>
<th>62</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>18</td>
<td>12</td>
<td>24</td>
<td>6</td>
<td>30</td>
<td>36</td>
</tr>
</tbody>
</table>

**Solution.** Here \( u = (x - 22)/4 \), \( v = (y - 24)/6 \), therefore

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>u</th>
<th>v</th>
<th>uv</th>
<th>u²</th>
<th>v²</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>18</td>
<td>-3</td>
<td>-1</td>
<td>3</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>12</td>
<td>-2</td>
<td>-2</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>18</td>
<td>24</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>22</td>
<td>6</td>
<td>0</td>
<td>-3</td>
<td>0</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>26</td>
<td>30</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>30</td>
<td>36</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>-3</td>
<td>-3</td>
<td>12</td>
<td>19</td>
<td>19</td>
</tr>
</tbody>
</table>

Therefore \( r = \frac{\sum uv - \frac{\sum u \sum v}{n}}{\sqrt{\left( \frac{\sum u^2}{n} - \left( \frac{\sum u}{n} \right)^2 \right) \left( \frac{\sum v^2}{n} - \left( \frac{\sum v}{n} \right)^2 \right)}} = \frac{12 - \frac{(-3)(-3)}{6}}{\sqrt{\left( \frac{19 - \frac{(-3)^2}{6}}{6} \right) \left( \frac{19 - \frac{(-3)^2}{6}}{6} \right)}} = 0.6. \)

Example 5. **Calculate the coefficient of correlation for the following table:**

| x/y | 0 - 5 | 5 - 10 | 10 - 15 | 15 - 20 | 20 - 25 |
Solution. Let \( A_x = 10 \) and \( A_y = 12.5 \) be the assumed values of \( x \) and \( y \) series \( h = 4, h' = 5 \) are their scales. Then

\[
\begin{align*}
\eta &= y - A_y \\
\xi &= x - A_x \\
\eta &= y - 12.5 \\
\xi &= x - 10 \\
u &= \frac{\eta}{5} \\
v &= \frac{\xi}{4}
\end{align*}
\]

<table>
<thead>
<tr>
<th>Class</th>
<th>Mid Value x</th>
<th>x - 10</th>
<th>( v = \eta/5 )</th>
<th>( u = \xi/4 )</th>
<th>0-5</th>
<th>5-10</th>
<th>10-15</th>
<th>15-20</th>
<th>20-25</th>
<th>Arrows Shows the Cell Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 4</td>
<td>2</td>
<td>-8</td>
<td>-2</td>
<td>-2</td>
<td>4</td>
<td>2</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>Total(f) fu fu' fuv</td>
</tr>
<tr>
<td>4 - 8</td>
<td>6</td>
<td>-4</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>8</td>
<td>17</td>
<td>-17 17 4</td>
</tr>
<tr>
<td>8 - 12</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>7   0 0</td>
</tr>
<tr>
<td>12 - 16</td>
<td>14</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3 4</td>
</tr>
</tbody>
</table>
Therefore the required correlation coefficient is

\[
r = \frac{\sum fuv - \frac{\sum fu \sum fv}{\sum f}}{\sqrt{\left(\sum fu^2 - \frac{(\sum fu)^2}{\sum f}\right) \left(\sum fv^2 - \frac{(\sum fv)^2}{\sum f}\right)}} = \frac{8 - (-20) \times 8}{30}
\]

\[
= \frac{8 + 5.333}{\sqrt{[32 - 13.33][28 - 2.13]}} = \frac{13.333}{21.977} = 0.6065 .
\]

Example 6. Find the coefficient of correlation for the following table:

<table>
<thead>
<tr>
<th>X →</th>
<th>67</th>
<th>72</th>
<th>77</th>
<th>82</th>
<th>87</th>
<th>92</th>
<th>97</th>
</tr>
</thead>
<tbody>
<tr>
<td>y ( \downarrow )</td>
<td>92</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>87</td>
<td>1</td>
<td>3</td>
<td>8</td>
<td>1</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>82</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>4</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>77</td>
<td>3</td>
<td>3</td>
<td>7</td>
<td>6</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>72</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>6</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>67</td>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>62</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Solution. Let \( u = (x - 82)/5 \) and \( v = (y - 77)/5 \). Then

<table>
<thead>
<tr>
<th>y \ u \ v</th>
<th>67 \ 72 \ 77 \ 82 \ 87 \ 92 \ 97</th>
<th>Total</th>
<th>fv</th>
<th>f v^2</th>
<th>f u v</th>
</tr>
</thead>
<tbody>
<tr>
<td>92 \ 3</td>
<td>0 \ 3 \ 6 \ 9 \ 7 \ 33</td>
<td>21 \ 63</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>87 \ 2</td>
<td>-2 \ 1 \ 2 \ 3 \ 4 \ 6 \ 18</td>
<td>36 \ 72</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>82 \ 1</td>
<td>-3 \ 4 \ 6 \ 9 \ 1 \ 28 \ 28</td>
<td></td>
<td></td>
<td></td>
<td>-15</td>
</tr>
<tr>
<td>77 \ 0</td>
<td>0 \ 3 \ 7 \ 6 \ 4 \ 23 \ 0</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>72 \ -1</td>
<td>3 \ 2 \ 1 \ 0 \ -1 \ -2 \ -18 \ 18</td>
<td></td>
<td></td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>67 \ -2</td>
<td>6 \ 4 \ 5 \ 6 \ -1 \ -2 \ -10 \ 20</td>
<td></td>
<td></td>
<td></td>
<td>25</td>
</tr>
<tr>
<td>62 \ -3</td>
<td>9 \ 1 \ -3 \ 9 \      \     \     \</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Therefore the required correlation coefficient is

\[
  r = -\frac{\sum f_{uv} - \sum f_{u} \sum f_{v}}{\sqrt{\left( \sum f_{u}^2 - \left( \sum f_{u} \right)^2 \right) \left( \sum f_{v}^2 - \left( \sum f_{v} \right)^2 \right)}} = \frac{115 - (-28) \times 54}{100} = \frac{115 - 1452}{100} = \frac{-1337}{100} = -13.37.
\]

Example 7. (a) If \( d_i \) stands for the difference in ranks of the \( i^{th} \) individual and and if \( d_i = 0 \) for all values of \( i \), then prove that \( r = 1 \).

(b) If \( d_i \) stands for the difference in ranks, then show that

(i) \( \min \Sigma d_i^2 = 0 \) (ii) \( \max \Sigma d_i^2 = (1/3)(n^3 - n) \)

(c) Show that in a ranked bivariate distribution in which no ties occurs and in which the variable are independent (i) \( \Sigma d_i^2 \) is always even

(ii) There are not more than \( (1/6)(n^3 - n) + 1 \) possible values of \( r \).

(d) If \( x_i + y_i = n + 1 \), then show that \( r = -1 \).

Solution. (a) Since \( d_i = 0 \) for all \( i \), therefore \( \Sigma d_i^2 = 0 \). Hence

\[
  r = 1 - \frac{6 \Sigma d^2}{n(n^2 - 1)} = 1 - 0 = 1.
\]
(b) Since the formula of rank correlation is \( r = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \). In case of perfect positive correlation \( r = 1 \), so that \( 1 = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \Rightarrow \sum d_i^2 = 0 \), which is possible only when \( d_i^2 = 0 \) for each \( i \) or if the ranks of both the variables are same. Similarly, in case of perfect negative correlation \( r = -1 \), so that \( -1 = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \Rightarrow \frac{6\sum d_i^2}{n(n^2 - 1)} = 2 \Rightarrow \sum d_i^2 = n(n^2 - 1)/3 = (n^3 - )/3 \). But since \( r \geq -1 \) so that \( 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \geq -1 \) which gives \( \frac{6\sum d_i^2}{n(n^2 - 1)} \leq 2 \). Hence \( \min \sum d_i^2 = 0 \) and \( \max \sum d_i^2 = (1/3)(n^3 - n) \).

(c) Let the ranks of two characteristics \( A \) and \( B \) of \( n \) individuals are \( x_1, x_2, x_3, \ldots, x_n \) and \( y_1, y_2, y_3, \ldots, y_n \) respectively. Then

\[
\Sigma x_i = 1 + 2 + 3 + \ldots + n = n(n + 1)/2 = \Sigma y_i \Rightarrow \bar{x} = \frac{n + 1}{2} = \bar{y} \quad \text{and} \\
\Sigma x_i^2 = 1^2 + 2^2 + 3^2 + \ldots + n^2 = n(n + 1)(2n + 1)/6 = \Sigma y_i^2 \Rightarrow \Sigma x_i^2/n = (n + 1)(2n + 1)/6 = \Sigma y_i^2/n.
\]

Therefore,

\[
\text{Var}(x) = \frac{1}{n} \sum (x_i - \bar{x})^2 = \frac{1}{n} \sum x_i^2 - \bar{x}^2 = \frac{(n + 1)(2n + 1)}{6} - \frac{(n + 1)^2}{4} = \frac{n^2 - 1}{12} = \text{Var}(y)
\]

Now, \( d_i = x_i - y_i = (x_i - \bar{x}) - (y_i - \bar{y}) \)

\[
\Rightarrow \sum d_i^2 = \sum (x_i - \bar{x})^2 + \sum (y_i - \bar{y})^2 - 2\sum (x_i - \bar{x})(y_i - \bar{y}) = n\text{Var}(x) + n\text{Var}(y) - 2n\text{Cov}(x, y) = 2n \frac{n^2 - 1}{12} - 2n\text{Cov}(x, y)
\]

\[
= 2\left\{n \frac{n^2 - 1}{12} - 2n\text{Cov}(x, y)\right\}, \quad \text{yields that } \Sigma d_i^2 \text{ is always even. Further since}
\]

\[
\max \Sigma d_i^2 = (1/3)(n^3 - n) \text{ which should be even, so the values of } \Sigma d_i^2 \text{ should be } 0, 2, 4, \ldots, (1/3)(n^3 - n) \text{ these are } (1/6)(n^3 - n) + 1 \text{ values.}
\]

(d) Sine \( d_i = x_i - y_i \) and \( x_i + y_i = n + 1 \), then \( d_i = 2x_i - (n + 1) \). Hence
\[
\sum d_i^2 = 4 \sum x_i^2 + \sum (n+1)^2 - 4(n+1) \sum x_i = 4 \frac{n(n+1)(2n+1)}{6} + n(n+1)^2 - 4(n+1) \frac{n(n+1)}{2} = \frac{1}{3} n(n^2 - 1).
\]
Hence \[ r = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} = 1 - \frac{1}{3} \frac{n(n^2 - 1)}{n(n^2 - 1)} = 1 - 2 = -1. \]

Example 8. Calculate the coefficient of correlation from the data given below by the method of differences:

<table>
<thead>
<tr>
<th>X:</th>
<th>78</th>
<th>89</th>
<th>97</th>
<th>69</th>
<th>59</th>
<th>79</th>
<th>68</th>
<th>57</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y:</td>
<td>125</td>
<td>137</td>
<td>156</td>
<td>112</td>
<td>107</td>
<td>136</td>
<td>123</td>
<td>108</td>
</tr>
</tbody>
</table>

Solution.

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Rank in X</th>
<th>Rank in Y</th>
<th>Rank diff. = X — Y = d</th>
<th>d^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>78</td>
<td>125</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>89</td>
<td>137</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>97</td>
<td>156</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>69</td>
<td>112</td>
<td>5</td>
<td>6</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>59</td>
<td>107</td>
<td>7</td>
<td>8</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>79</td>
<td>136</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>68</td>
<td>123</td>
<td>6</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>57</td>
<td>108</td>
<td>8</td>
<td>7</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>Σd = 0</td>
<td>Σd^2 = 4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Hence the required rank correlation coefficient is

\[ r = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 4}{8(64 - 1)} = 1 - \frac{3}{63} = \frac{20}{21} = 0.95. \]

Example 9. The ranking of ten students in two subjects A and B are as follows:
What is the coefficient of rank correlation?

Solution.

<table>
<thead>
<tr>
<th>Rank in A</th>
<th>Rank in B</th>
<th>Rank diff.</th>
<th>$d^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A - B = d$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>-3</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>-4</td>
<td>16</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>6</td>
<td>36</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>8</td>
<td>64</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>-9</td>
<td>81</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

$\Sigma d = 0$
$\Sigma d^2 = 214$

Hence the required rank correlation coefficient is

$$r = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 214}{10(100 - 1)} = \frac{-147}{495} = 0.3.$$ 

Example 10. Ten competitors in a beauty contest got marks by three judges in the following orders:

<table>
<thead>
<tr>
<th>First Judge</th>
<th>1</th>
<th>6</th>
<th>5</th>
<th>10</th>
<th>3</th>
<th>2</th>
<th>4</th>
<th>9</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Second Judge</td>
<td>3</td>
<td>5</td>
<td>8</td>
<td>4</td>
<td>7</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>Third Judge</td>
<td>6</td>
<td>4</td>
<td>9</td>
<td>8</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>10</td>
<td>5</td>
<td>7</td>
</tr>
</tbody>
</table>
Use the rank correlation coefficient to discuss which pair of judges have the nearest approach to common tastes in beauty.

Solution.

<table>
<thead>
<tr>
<th>First Judge</th>
<th>Second Judge</th>
<th>Third Judge</th>
<th>Rank diff.</th>
<th>Rank diff.</th>
<th>Rank diff.</th>
<th>$d_{12}^2$</th>
<th>$d_{13}^2$</th>
<th>$d_{23}^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>marks rank</td>
<td>Marks rank k</td>
<td>marks rank</td>
<td>$d_{12}$</td>
<td>$d_{13}$</td>
<td>$d_{23}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 10</td>
<td>3 8</td>
<td>6 5</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>25</td>
<td>9</td>
</tr>
<tr>
<td>6 5</td>
<td>5 6</td>
<td>4 7</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5 6</td>
<td>8 3</td>
<td>9 2</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>9</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>10 1</td>
<td>4 7</td>
<td>8 3</td>
<td>-6</td>
<td>-2</td>
<td>4</td>
<td>36</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>3 8</td>
<td>7 4</td>
<td>1 10</td>
<td>4</td>
<td>-2</td>
<td>-6</td>
<td>16</td>
<td>4</td>
<td>36</td>
</tr>
<tr>
<td>2 9</td>
<td>10 1</td>
<td>2 9</td>
<td>8</td>
<td>0</td>
<td>-8</td>
<td>64</td>
<td>0</td>
<td>64</td>
</tr>
<tr>
<td>4 7</td>
<td>2 9</td>
<td>3 8</td>
<td>-2</td>
<td>-1</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9 2</td>
<td>1 10</td>
<td>10 1</td>
<td>-8</td>
<td>1</td>
<td>9</td>
<td>64</td>
<td>1</td>
<td>81</td>
</tr>
<tr>
<td>7 4</td>
<td>6 5</td>
<td>5 6</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>8 3</td>
<td>9 2</td>
<td>2 4</td>
<td>1</td>
<td>-1</td>
<td>-2</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>$\Sigma d_{12}^2$ = 200</td>
<td>$\Sigma d_{13}^2$ = 60</td>
<td>$\Sigma d_{23}^2$ = 214</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Hence the required rank correlation coefficients is

$$r_{12} = 1 - \frac{6\sum d_{12}^2}{n(n^2 - 1)} = 1 - \frac{6 \times 200}{10(100 - 1)} = -\frac{7}{33},$$

$$r_{13} = 1 - \frac{6\sum d_{13}^2}{n(n^2 - 1)} = 1 - \frac{6 \times 60}{10(100 - 1)} = -\frac{7}{11},$$

$$r_{23} = 1 - \frac{6\sum d_{23}^2}{n(n^2 - 1)} = 1 - \frac{6 \times 214}{10(100 - 1)} = -\frac{49}{165}.$$ Here we conclude that first and third judges approach are nearest.

4.4.5. Regression: Line of Regression
Regression: By regression we mean that an association or relation between two variates \( x \), and \( y \). In case there exists association or relationship between two variates \( x \) and \( y \), and the points \((x_i, y_i)\) are values of \( x \) and \( y \) can be represented in \( xy \)-plane to give what is called a scatter or dot diagram. This diagram are useful in giving intuitive idea of the relationship between \( x \) and \( y \). Suppose that the scatter diagram indicates some relationship between the two variates \( x \) and \( y \), the dots of the scatter diagram will be more or less concentrated round a curve. This curve is called the curve of regression and the relationship is called curvilinear regression. In particular case when the curve is a straight line then it is called the line of regression and the regression is called linear. A line of regression is the straight line which gives the best fit in the least square sense to the given frequency. The method of least squares can be used to fit a straight line to the set of points given on the scatter diagram. Transfer the origin to the point \((M_x, M_y)\), where \( M_x \) and \( M_y \) are the means of \( x \)-series \( y \)-series respectively. Let \( x, y \) be the deviations from the respective means \( M_x \) and \( M_y \) i.e., \( x = X - M_x \) and \( y = Y - M_y \). Let \( Y = aX + b \) be the equation of the line of best fit of \( x \). Changing the origin to \((M_x, M_y)\) it will have the form \( y = ax + b \), where \( y = Y - M_y \) and \( x = X - M_x \).

Consider a dot \((x_r, y_r)\), then the difference between this point and the line is \( y_r - ax_r - b \). If \( U \), the sum of the squares of such distances i.e., \( U = \Sigma (y - ax - b)^2 \) for all \( r \). Now the principle of least squares, we choose \( a \) and \( b \) so that \( U \) is minimum by

\[
\frac{\partial U}{\partial a} = -2\Sigma x(y - ax - b) = 0 = \frac{\partial U}{\partial x} = -2\Sigma (y - ax - b), \quad \text{implies that } \Sigma xy - a\Sigma x^2 - b\Sigma x = 0 \quad \text{and} \quad \Sigma y - a\Sigma x - nb = 0.
\]

Since \( \Sigma x = 0 = \Sigma y \), we get \( a = \frac{\Sigma xy}{\Sigma x^2} = \frac{\sigma_y}{\sigma_x} \) and \( b = 0 \). Hence the line of fit is

\[
y = \frac{r\sigma_x}{\sigma_y} x \quad \text{or} \quad (Y - M_y) = \frac{r\sigma_x}{\sigma_y} (X - M_x).
\]

This is called the regression line of \( Y \) on \( X \). Thus if the straight line is so chosen that the sum and squares of deviations parallel to the axis of \( y \) is minimum, it is called the line of regression of \( y \) on \( x \). Similarly if \( X \) is taken to be dependent variable another regression line \( s(X - M_x) = \frac{r\sigma_x}{\sigma_y} (Y - M_y) \)

This is called regression line of \( X \) on \( Y \). Here \( \frac{r\sigma_y}{\sigma_x} \) and \( \frac{r\sigma_x}{\sigma_y} \) are called the regression coefficients of \( y \) on \( x \) and \( x \) on \( y \), denote by \( b_{yx} \) and \( b_{xy} \) respectively and \( b_{yx}b_{xy} \).
\[
\frac{r \sigma_x}{\sigma_x} \cdot \frac{r \sigma_y}{\sigma_y} = r^2 \quad \text{or the coefficient of correlation is the G. M. of the coefficients of regressions.}
\]

Note: (i) If \( r = +1 \) or \(-1 \), the two regression lines will coincide. The variables are perfectly correlated. If \( r = -1 \), the variables are perfectly negatively correlated, low values of one corresponding to high values of the other. If \( r = +1 \), variables are perfectly positively correlated, high values of one corresponding to high values of the other.

(ii) If \( r = 0 \), the two lines of regression become \( X = M_x \) and \( Y = M_y \), which are two lines parallel to \( Y \) and \( X \) axes respectively, passing through their means \( M_x \) and \( M_y \). They are perpendicular to each other. It means that mean values of \( X \) and \( Y \) do not change with \( y \) and \( X \) respectively i.e., \( X \) and \( Y \) are independent.

4.4.6. Covariance: If \( \bar{x} - \bar{y} \) be the expected values (or means) of two variates \( x \) and \( y \), then the covariance between \( x \) and \( y \) is defined by the relation

\[
\text{Cov}(x, y) = E(x - \bar{x})(y - \bar{y}) = E\{E(x)\{y - E(y)\} = E[xy - xE(x) - yE(y) + E(x)E(y)]
\]

\[
\]

**Note:** The covariance of two independent variates is equal to zero. Since if \( x \) and \( y \) are independent, then \( \text{Cov}(x, y) = E(x - \bar{x})(y - \bar{y}) = E(x)E(y) = 0 \), since here \( E(x - \bar{x}) = E(x) - \bar{x} = 0 \) = \( E(y - \bar{y}) \).

Correlation coefficient: With the above notation the correlation coefficient \( r \) is defined by the formula:

\[
r = \frac{\text{Cov}(x, y)}{\sqrt{E(x - \bar{x})^2E(y - \bar{y})}} = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x)\text{Var}(y)}} = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y}.
\]

If \( x \) and \( y \) are independent variates then we have \( \text{cov}(x, y) = 0 \), then \( r = 0 \) i.e., they are uncorrelated.

Theorem 1(Schwarz’s Inequality): For any stochastic variates \( x \) and \( y \)

\[
[E(xy)]^2 \leq E(x^2)E(y^2).
\]

**Proof.** For any real constant \( a \), \( (ax - y)^2 \geq 0 \), therefore \( E(ax - y)^2 \geq 0 \),

i.e., \( a^2E(x^2) + E(y^2) - aE(xy) \geq 0 \). Since a being arbitrary so put \( a = \frac{E(xy)}{E(x^2)} \), then

\[
\left[\frac{E(xy)}{E(x^2)}\right]^2 E(x^2) + E(y^2) - \frac{E(xy)}{E(x^2)} E(xy) \geq 0 \Rightarrow [E(xy)]^2 + E(x^2)E(y^2) - 2[E(xy)]^2 \geq 0
\]
i.e., \[ (E(xy))^2 \leq E(x^2)E(y^2). \]

Theorem 2: Sign of correlation coefficient and regression coefficients are same.

Proof. Since \( b_{yx} = r_{yx} \frac{\sigma_y}{\sigma_x} = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} \), similarly, \( b_{xy} = \frac{\text{cov}(x, y)}{\sigma_x^2} \) and \( r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} \), since \( \sigma_x, \sigma_y \) are always positive. Hence sign of \( b_{yx}, b_{xy} \) and \( r \) depend on the sign of \( \text{Cov}(x, y) \).

Theorem 3: If a regression coefficient is greater than 1 then the other is less than 1.

Proof. Since if \( b_{yx} \geq 1 \) then \( 1/b_{yx} \leq 1 \). Now since \( b_{yx} b_{xy} = r^2 \leq 1 \) implies that \( b_{xy} \leq 1 \).

4.4.7. Illustrative Examples

Example 1. Prove that the Pearson's coefficient of correlation \( r \) lies between -1 and +1.

Solution. Let \( U = \Sigma(y - ax - b)^2 \), then \( U \) is minimum if

\[
\frac{\partial U}{\partial a} = -2\Sigma x(y - ax - b) = 0 = \frac{\partial U}{\partial x} = -2\Sigma (y - ax - b),
\]

implying that \( \Sigma xy - a\Sigma x^2 - b\Sigma x = 0 \) and \( \Sigma y - a\Sigma x - nb = 0 \). Since \( \Sigma x = 0 = \Sigma y \), we get \( a = \frac{\Sigma xy}{\Sigma x^2} = \frac{r \sigma_y}{\sigma_x} \) and \( b = 0 \). Therefore

\[
U = \Sigma(y - ax)^2 = \Sigma y^2 - 2a\Sigma xy + a^2\Sigma x^2 = \Sigma y^2 - \frac{\Sigma xy}{\Sigma x^2} \Sigma x^2 - 2 \left( \frac{\Sigma xy}{\Sigma x^2} \right)^2 \Sigma x^2 + \left( \frac{\Sigma xy}{\Sigma x^2} \right)^2 \Sigma x^2
\]

\[
= \Sigma y^2 - \left( \frac{\Sigma xy}{\Sigma x^2} \right)^2 \Sigma x^2 = \Sigma y^2 \left[ 1 - \left( \frac{\Sigma xy}{\Sigma x^2} \right)^2 \right] = \Sigma y^2 (1 - r^2) \geq 0,
\]

because \( U \) is the sum of the squares will not be negative. Also \( \Sigma y^2 \) will not be negative. Hence \( 1 - r^2 \geq 1 \) implies that limit of \( r \) as -1 and +1.

Example 2. Prove that A. M. of the coefficient of regression is greater than the coefficient of correlation.

\[
\frac{r \sigma_y}{\sigma_x} + \frac{r \sigma_x}{\sigma_y} > r
\]

Solution. Here we are to prove that \( \frac{r \sigma_y}{\sigma_x} + \frac{r \sigma_x}{\sigma_y} > r \)

or \( \sigma_x^2 + \sigma_y^2 > 2\sigma_x \sigma_y \) or \( (\sigma_x - \sigma_y)^2 > 0 \), which is true.
Example 3. If $\theta$ is the acute angle between the two regression lines, in the case of two variables $x$ and $y$, show that

$$\tan \theta = \frac{1 - r^2}{r} \cdot \frac{\sigma_y}{\sigma_x^2 + \sigma_y^2},$$

where $r$, $\sigma_x$, $\sigma_y$ have their usual meanings. Moreover, explain the significance of the formula when $r = 0$ and $r = \pm 1$.

Solution. If $\theta_1$, $\theta_2$ are the angles which the two regression lines make with the $x$-axis, then

$$\tan \theta = \tan(\theta_1 - \theta_2) = \frac{\frac{\sigma_y}{\sigma_x} - \frac{r \sigma_y}{\sigma_x}}{1 + \frac{r \sigma_y}{\sigma_x}} = \frac{1 - r^2}{r} \cdot \frac{\sigma_y}{\sigma_x^2 + \sigma_y^2}.$$

When $r = 0$ then $\theta = \pi/2$ i.e., the two lines of regression are perpendicular to each other.

The estimated value of $y$ is the same for all values of $x$ or vice-versa. Also if $r = \pm 1$ then $\theta = 0$, hence the lines of regression coincide and there is a perfect correlation between the two variates $x$ and $y$.

Example 4. The following marks have been obtained by a class of students in Statistics.

<table>
<thead>
<tr>
<th>Paper I</th>
<th>45</th>
<th>55</th>
<th>56</th>
<th>58</th>
<th>60</th>
<th>65</th>
<th>68</th>
<th>70</th>
<th>75</th>
<th>80</th>
<th>85</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper II</td>
<td>56</td>
<td>50</td>
<td>48</td>
<td>60</td>
<td>62</td>
<td>64</td>
<td>65</td>
<td>70</td>
<td>74</td>
<td>82</td>
<td>90</td>
</tr>
</tbody>
</table>

Compute the coefficient of correlation for the above data. Find also the equation of the lines of regression.

Solution. Taking assume means 65 and 70 respectively of first paper ($x$) and second paper ($y$) respectively, we get the following table:

<table>
<thead>
<tr>
<th>Paper I</th>
<th>$x$= $x - 65$</th>
<th>$x^2$</th>
<th>Paper II</th>
<th>$y$</th>
<th>$y - 70$</th>
<th>$y^2$</th>
<th>$\xi \eta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td></td>
<td></td>
<td>y</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>-20</td>
<td>400</td>
<td>56</td>
<td>-14</td>
<td>196</td>
<td>280</td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>-10</td>
<td>100</td>
<td>50</td>
<td>-20</td>
<td>400</td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>-9</td>
<td>81</td>
<td>48</td>
<td>-22</td>
<td>484</td>
<td>198</td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>-7</td>
<td>49</td>
<td>60</td>
<td>-10</td>
<td>100</td>
<td>70</td>
<td></td>
</tr>
</tbody>
</table>
The table shows the heights of fathers and sons in inches:

<table>
<thead>
<tr>
<th></th>
<th>-5</th>
<th>25</th>
<th>62</th>
<th>-8</th>
<th>64</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>0</td>
<td>0</td>
<td>64</td>
<td>-6</td>
<td>36</td>
<td>0</td>
</tr>
<tr>
<td>65</td>
<td>3</td>
<td>9</td>
<td>65</td>
<td>-5</td>
<td>25</td>
<td>-15</td>
</tr>
<tr>
<td>70</td>
<td>5</td>
<td>25</td>
<td>70</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>75</td>
<td>10</td>
<td>100</td>
<td>74</td>
<td>4</td>
<td>16</td>
<td>40</td>
</tr>
<tr>
<td>80</td>
<td>15</td>
<td>225</td>
<td>82</td>
<td>12</td>
<td>144</td>
<td>180</td>
</tr>
<tr>
<td>85</td>
<td>20</td>
<td>400</td>
<td>90</td>
<td>20</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td>Total</td>
<td>2</td>
<td>1414</td>
<td>11</td>
<td>1865</td>
<td>1393</td>
<td></td>
</tr>
</tbody>
</table>

Then \( r = \frac{\sum \xi \eta - \frac{\sum \xi \sum \eta}{n}}{\sqrt{\left( \sum \xi^2 - \frac{(\sum \xi)^2}{n} \right) \left( \sum \eta^2 - \frac{(\sum \eta)^2}{n} \right)}} = \frac{1393 - \frac{2(-49)}{11}}{\sqrt{\left( \frac{1414 - 4}{11} \right) \left( \frac{1865 - 49^2}{11} \right)}} = 0.918 \)

Here \( n = 11 \), now \( \sigma_x = \sqrt{\frac{\sum \xi^2 - (\sum \xi)^2}{n}} \), \( \sigma_y = \sqrt{\frac{\sum \eta^2 - (\sum \eta)^2}{n}} \)

\[ \Rightarrow \sigma_x = \sqrt{\frac{1414 - \left( \frac{2}{11} \right)^2}{11}} = 11.336, \sigma_y = \sqrt{\frac{1865 - \left( \frac{-49}{11} \right)^2}{11}} = 12.235 \]

Therefore regression coefficient of \( y \) on \( x \), \( b_{yx} = \frac{r \sigma_y}{\sigma_x} = \frac{0.919 \times 12.235}{11.336} = 0.992 \) and regression coefficient of \( x \) on \( y \), \( b_{xy} = \frac{r \sigma_x}{\sigma_y} = \frac{0.919 \times 11.336}{12.235} = 0.851 \). Also means of \( x \)-series, \( M_x = \) Assumed mean + \( \frac{\sum \xi}{n} = 65 + \frac{2}{11} = 65.2 \) and

\[ y \text{-series, } M_y = \text{Assumed mean} + \frac{\sum \eta}{n} = 70 + \frac{-49}{11} = 65.55 \]

Hence the required equations of regression lines:

\[ y \text{ on } x \text{ is } y - 65.55 = 0.99(x - 65.2) \text{ and } x \text{ on } y \text{ is } (x - 65.2) = 0.85(y - 65.55). \]

Example 5. *Heights of fathers and sons are given in inches:*

<table>
<thead>
<tr>
<th>60</th>
<th>-5</th>
<th>25</th>
<th>62</th>
<th>-8</th>
<th>64</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>65</td>
<td>0</td>
<td>0</td>
<td>64</td>
<td>-6</td>
<td>36</td>
<td>0</td>
</tr>
<tr>
<td>68</td>
<td>3</td>
<td>9</td>
<td>65</td>
<td>-5</td>
<td>25</td>
<td>-15</td>
</tr>
<tr>
<td>70</td>
<td>5</td>
<td>25</td>
<td>70</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>75</td>
<td>10</td>
<td>100</td>
<td>74</td>
<td>4</td>
<td>16</td>
<td>40</td>
</tr>
<tr>
<td>80</td>
<td>15</td>
<td>225</td>
<td>82</td>
<td>12</td>
<td>144</td>
<td>180</td>
</tr>
<tr>
<td>85</td>
<td>20</td>
<td>400</td>
<td>90</td>
<td>20</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td>Total</td>
<td>2</td>
<td>1414</td>
<td>11</td>
<td>1865</td>
<td>1393</td>
<td></td>
</tr>
</tbody>
</table>
Form the two lines of regression and calculate the expected average height of the son when the height of the father is 67.5 inches.

Solution. Here the table of calculation is

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
<th>$\xi = x - 69$</th>
<th>$\eta = y - 69$</th>
<th>$\xi^2$</th>
<th>$\eta^2$</th>
<th>$\xi \eta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>65</td>
<td>67</td>
<td>-4</td>
<td>-2</td>
<td>16</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>66</td>
<td>68</td>
<td>-3</td>
<td>-1</td>
<td>9</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>67</td>
<td>64</td>
<td>-2</td>
<td>-5</td>
<td>4</td>
<td>25</td>
<td>10</td>
</tr>
<tr>
<td>67</td>
<td>68</td>
<td>-2</td>
<td>-1</td>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>68</td>
<td>72</td>
<td>-1</td>
<td>3</td>
<td>1</td>
<td>9</td>
<td>-3</td>
</tr>
<tr>
<td>69</td>
<td>70</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>71</td>
<td>69</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>73</td>
<td>70</td>
<td>4</td>
<td>1</td>
<td>16</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-6</td>
<td>-4</td>
<td>54</td>
<td>42</td>
<td>24</td>
</tr>
</tbody>
</table>

Here $n = 8$, means of $x$-series, $M_x = \text{Assumed mean} + \frac{\sum \xi}{n} = 69 + \frac{-6}{8} = 68.25$ and $y$-series, $M_y = \text{Assumed mean} + \frac{\sum \eta}{n} = 69 + \frac{-4}{8} = 68.5$. Further,

$$ \sigma_x = \sqrt{\frac{\sum \xi^2}{n} - \left( \frac{\sum \xi}{n} \right)^2}$$

$$ \Rightarrow \sigma_x = \sqrt{\frac{54}{8} - \left( \frac{-6}{8} \right)^2} = 2.49$$

$$ \sigma_y = \sqrt{\frac{42}{8} - \left( \frac{-4}{8} \right)^2} = 2.23$$

and
\[
\begin{align*}
    r &= \frac{\sum \xi \eta - \frac{\sum \xi \sum \eta}{n}}{\sqrt{\left(\sum \xi^2 - \frac{(\sum \xi)^2}{n}\right)\left(\sum \eta^2 - \frac{(\sum \eta)^2}{n}\right)}} = \frac{24 - 3}{\sqrt{\left\{54 - \frac{9}{2}\right\}\left\{42 - 2\right\}}} = 0.47.
\end{align*}
\]

Therefore regression coefficient of \( y \) on \( x \), \( b_{yx} = \frac{r \sigma_y}{\sigma_x} = \frac{0.47 \times 2.23}{2.49} = 0.421 \) and regression coefficient of \( x \) on \( y \), \( b_{xy} = \frac{r \sigma_x}{\sigma_y} = \frac{0.47 \times 2.49}{2.23} = 0.52 \). Hence the lines of regression are:

\[y \text{ on } x \quad y = 68.5 = 0.421(x - 68.25) \quad \text{and} \quad x \text{ on } y \quad (x - 68.25) = 0.52(y - 68.5).\]

Now for the height of son \( x = 67.5 \), the corresponding height of father is given by \( y = 68.19 \).

Example 6. The following data are available in a factory: Variance of \( x \) = 9, Regression equations: \( 8x - 10y + 66 = 0, 40x - 18y = 214 \). Calculate (a) the mean values of \( x \) and \( y \) (b) the standard deviation of \( y \), and (c) the coefficient of correlation between \( x \) and \( y \).

Solution. On solving the given regression equation, we get the means of \( x \) and \( y \) are: \( x = 13, \ y = 17 \). Also the given regression equations can be put as:

\[y = 0.8x + 6.6; \ x = 0.45y + 5.35,\]

we obtain the regression coefficients:

\[y \text{ on } x \quad b_{yx} = 0.8 \quad \text{and} \quad x \text{ on } y \quad b_{xy} = 0.25 \] and the correlation coefficient is given by

\[r^2 = b_{yx} b_{xy} = 0.8 \times 0.45 = 0.360 \quad \text{gives} \quad r = 0.6. \] Further variance of \( x \), \( \sigma_x^2 = 9 \), hence \( \sigma_x = 3 \). Therefore \( b_{yx} = r \sigma_y / \sigma_x = 0.6 \sigma_y / 3 = 0.8 \) gives \( \sigma_y = 4 \).

Example 7. If \( x \) and \( y \) are two uncorrelated variables and if \( u = x + y, v = x - y \), find the correlation between \( u \) and \( v \) in terms of \( \sigma_x \) and \( \sigma_y \), the s.d.'s of \( x \) and \( y \) respectively.

Solution. In terms of \( u \) and \( v \), \( r \) the coefficient of correlation is given by

\[r = \frac{\text{Cov}(u, v)}{\sqrt{\text{Var}(u) \cdot \text{Var}(v)}}. \]

Here \( u = x + y \) and \( v = x - y \) so that \( \bar{u} = \bar{x} + \bar{y}, \bar{v} = \bar{x} - \bar{y} \). Then

\[\text{Cov}(u, v) = E(u - \bar{u})(v - \bar{v}) = E((x - \bar{x} + y - \bar{y})(x - \bar{x} - y + \bar{y})) = E[(x - \bar{x})^2 - (y - \bar{y})^2] \]

\[= E(x - \bar{x})^2 - E(y - \bar{y})^2 = \sigma_x^2 - \sigma_y^2. \]

Also
\( \text{Var}(u) = \text{Var}(x + y) = E((x + y)^2) = E(x^2) + E(y^2) + 2E(xy) = \text{Var}(x) + \text{Var}(y) + 2\text{Cov}(x, y) \)

Similarly, \( \text{Var}(v) = \text{Var}(x - y) = \text{Var}(x) + \text{Var}(y) - 2\text{Cov}(x, y) \). Now \( \text{Cov}(x, y) = 0 \) as \( x \) and \( y \) are uncorrelated. Hence

\( \text{Var}(u) = \text{Var}(x) + \text{Var}(y) = \text{Var}(v) \Rightarrow \sigma_u = \sqrt{\sigma_x^2 + \sigma_y^2} = \sigma_y. \)

Therefore

\[
\rho = \frac{\text{cov}(u, v)}{\sqrt{\text{var}(u) \text{var}(v)}} = \frac{\sigma_x^2 - \sigma_y^2}{\sigma_x^2 + \sigma_y^2}. 
\]

Example 8. Two independent variables \( x \) and \( y \) have means 5 and 10, variances 4 and 9 respectively. Obtain in the coefficient of correlation between \( u \) and \( v \) where

\[ u = 3x + 4y \quad \text{and} \quad v = 3x - y. \]

Solution. Here \( E(u) = 3E(x) + 4E(y), E(v) = 3E(x) - E(y). \)

Hence \( u - E(u) = 3[x - E(x)] + 4[y - E(y)] \) and \( v - E(v) = 3[x - E(x)] - [y - E(y)]. \) Therefore \( \text{Cov}(u, v) = E(u - \overline{u})(v - \overline{v}) = 9E[(x - \overline{x})^2] - 4E[(y - \overline{y})^2] + 9E[(x - \overline{x})(y - \overline{y})] \)

\[ \text{Cov}(u, v) = 9 \text{var}(x) - 4 \text{var}(y) = 9 \times 4 - 4 \times 9 = 0 \Rightarrow r = \frac{\text{cov}(u, v)}{\sigma_u \sigma_v} = 0. \]

Example 9. Show that the coefficient of correlation between two variables \( x \) and \( y \) is given by

\[ r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x,y}^2}{2\sigma_x \sigma_y}. \]

Solution. Since \( \sigma_{x,y}^2 = E(x - y)^2 = E(x^2) + E(y^2) - 2E(x, y) = \sigma_x^2 + \sigma_y^2 - 2 \text{cov}(x, y) \)

\[ \Rightarrow \text{cov}(x, y) = \frac{1}{2}(\sigma_x^2 + \sigma_y^2 - \sigma_{x,y}^2) \Rightarrow r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} = \frac{\sigma_y^2 - \sigma_{x,y}^2}{2\sigma_x \sigma_y}. \]

Example 10. Find the correlation coefficient between \( x \) and \( a - x. \)

Solution. Let \( u = a - x \) then \( \overline{u} = a - \overline{x} \Rightarrow \text{var}(u) = E((u - \overline{u})^2) = E(x - \overline{x})^2 = \sigma_u \)

\[ \text{Cov}(x, u) = E((x - \overline{x})(u - \overline{u})) = -E(x - \overline{x})^2 = -\sigma_u^2 \Rightarrow r = \frac{\sigma_u^2}{\sigma_u \sigma_u} = -1. \]

Example 11. If \( x_1, x_2, x_3 \) are three uncorrelated variates having standard deviations \( \sigma_1, \sigma_2, \sigma_3 \) respectively. Obtain the coefficients of correlation between \( (x_1 + x_2) \) and \( x_2 + x_3). \)
Solution. Let \( u = x_1 + x_2 \) and \( v = x_2 + x_3 \), then \( E(u) = E(x_1) + E(x_2) \) and \( E(v) = E(x_2) + E(x_3) \) implies that \( u - E(u) = x_1 - E(x_1) + x_2 - E(x_2) \).

Squaring both sides and taking expected values, we get

\[
E\{u - E(u)\}^2 = E\{x_1 - E(x_1)\}^2 + 2E\{x_1 - E(x_1)\} E\{x_2 - E(x_2)\} + 2E\{x_2 - E(x_2)\}^2
\]

\[\Rightarrow \text{var}(u) = \text{var}(x_1) + \text{var}(x_2) + 2\text{cov}(x_1, x_2) = \sigma_1^2 + \sigma_2^2 \text{ as } \text{cov}(x_1, x_2) = 0.\]

Similarly, \( \text{var}(v) = \sigma_2^2 + \sigma_3^2 \). Again

\[
\text{Cov}(u, v) = E\{(u - \bar{u})(v - \bar{v})\} = E\{(x_1 - \bar{x}_1 + x_2 - \bar{x}_2)(x_2 - \bar{x}_3 + x_3 - \bar{x}_3)\}
\]

\[= E\{(x_1 - \bar{x}_1)(x_2 - \bar{x}_2) + (x_2 - \bar{x}_2)(x_3 - \bar{x}_3) + (x_2 - \bar{x}_3)^2\}
\]

\[= E(x_2 - \bar{x}_2)^2 + \text{cov}(x_1, x_2) + \text{cov}(x_2, x_3) = \sigma_2^2. \text{ Hence}
\]

\[r = \frac{\text{cov}(u, v)}{\sqrt{\text{var}(u) \text{var}(v)}} = \frac{\sigma_2^2}{\sqrt{[(\sigma_1^2 + \sigma_2^2)(\sigma_2^2 + \sigma_3^2)]}}.
\]

Example 12. Correlation coefficient between two variables x and y is 0.32, covariance is 7.86 and variance of x is 10. Find the variance of y.

Solution. We have given that \( r = 0.32 \), \( \text{Cov}(x, y) = 7.86 \) and \( \sigma_x = 10 \). Now

\[r = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y} = \sigma_y = \frac{\text{Cov}(x, y)}{r \sigma_x} = \frac{7.86}{0.32 \times 10} = 2.45625 . \text{ Therefore Var}(y) = 6 \text{ appr.}
\]

4.5. Partial and Multiple Correlation: In our previous sections we have discussed the correlation between two variates only. When the values of one variable are influenced by those of another, the coefficient of correlation provides a useful measure of the degree of association between them. But it often happens that the values of a variable are influenced by those of several others. Therefore it becomes necessary to find correlation between three or more variates. For examples

(i) Number of children \( (x_1) \) per family depending on income \( (x_2) \) of the family and age \( (x_3) \) at marriage.

(ii) Agricultural production \( (x_1) \) depends on rainfall \( (x_2) \) and on fertilizers \( (x_3) \).

(iii) Crimes \( (x_1) \) depending upon illiteracy \( (x_2) \) and increased population \( (x_3) \).
Here we shall consider the case of three mutually correlated variables only.

When we consider the combined influence of two or more variates \((x_1, x_2, x_3, \ldots)\) upon a variate not included in the above variates our study is of multiple correlation. The degree of relationship existing between three or more variables is called multiple correlation. The correlation between two variates \((x_1, x_2)\) when the linear effect of the third variate \((x_3)\) in them has been eliminated from both is called partial correlation. For example if we study the effects of both rainfall \((x_2)\) and fertilizers \((x_3)\) on the production of whets then it leads an example of multiple correlation, but if we eliminate the effect of any one of \((x_2)\) or \((x_3)\) then the correlation between \((x_3, x_1)\) or \((x_2, x_1)\) is an example of partial correlation.

4.5.1. Distribution of Variables.

Two variables. Let \(\sigma_1\) and \(\sigma_2\) be the standard deviations of two variables \(x_1\) and \(x_2\) measured from their means (i.e., \(x_1 = X_1 - \bar{x}_1\) and \(x_2 = X_2 - \bar{x}_2\)), where \(E(x_1) = 0 = E(x_2)\). Let the lines of regression of \(x_1\) on \(x_2\) and of \(x_2\) on \(x_1\) be denoted by the symbols:

\[
x_1 = b_{12}x_2 \quad \text{and} \quad x_2 = b_{21}x_1, \quad \text{here} \quad b_{12} \quad \text{and} \quad b_{21} \quad \text{are the regression coefficient of the line of regressions} \quad x_1 \quad \text{on} \quad x_2 \quad \text{and} \quad x_2 \quad \text{on} \quad x_1 \quad \text{respectively.}
\]

Denote \(x_{1.2} = x_1 - b_{12}x_2\) and \(x_{2.1} = x_2 - b_{21}x_1\), here \(b_{12}\) and \(b_{21}\) are the regression coefficient of the line of regressions \(x_1\) on \(x_2\) and \(x_2\) on \(x_1\) respectively.

Here \(x_{1.2}\) and \(x_{2.1}\) are called residuals which are the deviations of the representatives points from the corresponding line of regression. The values of \(b_{12}\) and \(b_{21}\) are obtained by the principle of least squares which leads to normal equations:

\[
\Sigma(x_1 - b_{12}x_2) = 0 = \Sigma x_2(x_1 - b_{12}x_2) \quad \text{and} \quad \Sigma(x_2 - b_{21}x_1) = 0 = \Sigma x_1(x_2 - b_{21}x_1).
\]

In another notation these normal equations can be expressed as:

\[
\Sigma x_{1.2} = 0 = \Sigma x_2 x_{1.2} \quad \text{and} \quad \Sigma x_{2.1} = 0 = \Sigma x_1 x_{2.1}.
\]

These equations yields that: \(b_{12} = \frac{\Sigma x_1 x_2}{\Sigma x_2^2}\) and \(b_{21} = \frac{\Sigma x_1 x_2}{\Sigma x_1^2}\). Let the coefficients of correlation between \(x_1\) and \(x_2\) be denoted by \(r_{12}\) or \(r_{21}\). Then

\[
r_{12}^2 = b_{12}b_{21} \quad \text{and} \quad b_{12} = r_{12} \frac{\sigma_1}{\sigma_2}, \quad b_{21} = r_{21} \frac{\sigma_2}{\sigma_1}.
\]

If we denote the mean squares of deviations of \((i)\) by \(\sigma_{1.2}^2\) and \(\sigma_{2.1}^2\), then
\[ \sigma_{1,2}^2 = \frac{1}{N} \sum x_{1,2}^2 = \sigma_1^2 (1 - r_{12}^2) \] and \[ \sigma_{2,1}^2 = \frac{1}{N} \sum x_{2,1}^2 = \sigma_2^2 (1 - r_{21}^2). \]

Three variables. Let \( \sigma_1, \sigma_2 \) and \( \sigma_3 \) be the standard deviations of three variables \( x_1, x_2 \) and \( x_3 \) measured from their means (i.e., \( x_1 = X_1 - \bar{X}_1, \ x_2 = X_2 - \bar{X}_2 \) and \( x_3 = X_3 - \bar{X}_3 \)), where \( E(x_1) = 0 = E(x_2) \). Let the plane of regression of \( x_1 \) depending on \( x_2 \) and \( x_3 \) is:

\[ x_1 = a + b_{12,3} x_2 + b_{13,2} x_3, \] here \( b_{12,3} \) and \( b_{13,2} \) are the slopes of straight lines in the graph of \( x_1, x_2 \) keeping \( x_3 \) as constant and \( x_1, x_3 \) keeping \( x_2 \) as constant respectively. These are known as the partial regression coefficients. Here the first subscript attached to the b’s is the subscript of the dependent variable for which estimate is being found, the second subscript is that of variable x which the coefficient multiplies. These are called the primary subscripts. After the primary subscripts, and separated from them by a dot, are placed the subscripts of the other variables that enter into that equation. These are called the secondary subscripts and their number determines the order of the regression coefficients. The values of a’s and b’s are obtained by the principle of least squares.

Let \( U = \Sigma (x_1 - a - b_{12,3} x_2 - b_{13,2} x_3) = \Sigma x_{1,23}^2 \). Here \( x_{1,23} \) called residuals which are the deviations of the representatives points from the corresponding plane regression.

Now \( \frac{\partial U}{\partial a} = \frac{\partial U}{\partial b_{12,3}} = \frac{\partial U}{\partial b_{13,2}} = 0 \), gives the normal equations are:

\[ \Sigma (x_1 - a - b_{12,3} x_2 - b_{13,2} x_3) = 0 = \Sigma x_{1,23} \] ... ... ... (i)

\[ \Sigma x_2 (x_1 - a - b_{12,3} x_2 - b_{13,2} x_3) = 0 = \Sigma x_{2,23} x_{1,23} \] ... ... ... (ii)

\[ \Sigma x_3 (x_1 - a - b_{12,3} x_2 - b_{13,2} x_3) = 0 = \Sigma x_{3,23} x_{1,23} \] ... ... ... (iii).

Now since \( \Sigma x_1 = \Sigma x_2 = \Sigma x_3 = 0 \), then by (i) \( a = 0 \) and (ii), (iii) reduced to

\[ b_{12,3} \Sigma x_2^2 + b_{13,2} \Sigma x_2 x_3 - \Sigma x_1 x_2 = 0 \] ... ... ... (iv)

\[ b_{12,3} \Sigma x_2 x_3 + b_{13,2} \Sigma x_3^2 - \Sigma x_1 x_3 = 0 \] ... ... ... (v).

Since \( \Sigma x_2^2 = N\sigma_2^2, \Sigma x_3^2 = N\sigma_3^2, \Sigma x_2 x_3 = N r_{23} \sigma_2 \sigma_3 \) etc., where \( r_{ij} \) is coefficient of correlation between variables \( x_i \) and \( x_j \), therefore from equations (iv) and (v), we get

\[ b_{12,3} \sigma_2^2 + b_{13,2} (r_{23} \sigma_2 \sigma_3) - r_{12} \sigma_1 \sigma_2 = 0 \] or \[ b_{12,3} \sigma_2 + b_{13,2} r_{23} \sigma_3 = r_{12} \sigma_1 \] ... (vi)

\[ b_{12,3} (r_{23} \sigma_2 \sigma_3) + b_{13,2} \sigma_3^2 - r_{13} \sigma_1 \sigma_3 = 0 \] or \[ b_{12,3} r_{23} \sigma_2 + b_{13,2} \sigma_3 = r_{13} \sigma_1 \] ... (vii).

On solving equations (vi) and (vii) for \( b_{12,3} \) and \( b_{13,2} \), we have
\[ b_{123} = \frac{\sigma_1}{\sigma_2} \left( \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} \right) = \frac{\sigma_1}{\sigma_2} \begin{vmatrix} r_{12} & r_{23} \\ r_{13} & 1 \\ r_{23} & 1 \end{vmatrix}, \quad b_{132} = \frac{\sigma_1}{\sigma_3} \left( \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} \right) = \frac{\sigma_1}{\sigma_3} \begin{vmatrix} r_{13} & r_{23} \\ r_{12} & 1 \\ r_{23} & 1 \end{vmatrix} \] etc.

4.5.2. Properties of Residuals.

Property 1. The sum of the products of corresponding values of a variate and a residual is zero provided the subscript of the variate occurs among the secondary subscripts of the residual.

Proof. Let the regression equation of \( x_1 \) on \( x_2 \) and \( x_3 \) be \( x_1 = b_{123}x_2 + b_{132}x_3 \) then the normal equations are \( \Sigma x_2 x_{123} = 0 = \Sigma x_3 x_{1,23} \) etc.

Property 2. The sum of the products of two residuals is unaltered by omitting from the residual any or all of the secondary subscripts which are common to both.

Proof. Let \( x_{1,2} = x_1 - b_{12}x_2 \), then with the help of the normal equations
\[ \Sigma x_{1,23}x_{1,2} = \Sigma x_{1,23}(x_1 - b_{12}x_2) = \Sigma x_{1,23}x_1, \quad \Sigma x_{1,23}x_{1,23} = \Sigma x_{1,23}(x_1 - b_{123}x_2 - b_{132}x_3) = \Sigma x_{1,23}x_1 \]

Property 3. The sum of the products of two residuals is zero provided all the subscripts of one residual occur among the secondary subscripts of the second.

Proof. From normal equations, we get
\[ \Sigma x_{3,2}x_{1,23} = \Sigma(x_3 - b_{32}x_2)x_{1,23} = 0, \text{ similarly } \Sigma x_{2,3}x_{1,23} = 0. \]

4.5.3. Multiple and Partial correlation Coefficients:

**Standard Deviation of the Residuals.** The standard deviation of \( x_1 \) and \( x_2 \), keeping \( x_3 \) constant is denoted by \( \sigma_{1,23} \), since \( a = 0 \), we have
\[
\sigma_{1,23}^2 = \frac{1}{N} \sum x_{1,23}^2 = \frac{1}{N} \sum (x_1 - b_{123}x_2 - b_{132}x_3)^2
\]
\[
\sum(x_i - b_{123}x_2 - b_{132}x_3)(x_i - b_{123}x_2 - b_{132}x_3) = \frac{1}{N} \sum x_i(x_i - b_{123}x_2 - b_{132}x_3)
\]

\[
= \sigma_i^2 - b_{123}\sigma_i\sigma_{12}r_{12} - b_{132}\sigma_i\sigma_{13}r_{13} = \sigma_i^2 - \frac{r_{12}r_{13}r_{23}}{1 - r_{23}^2}\sigma_i^2 - \frac{r_{12}r_{13}r_{23}}{1 - r_{23}^2}\sigma_i^2 r_{13}
\]

\[
= -\frac{\sigma_i^2}{1 - r_{23}^2}(1 - r_{12}^2 - r_{13}^2 + 2r_{12}r_{13}r_{23})
\]

\[
= \sigma_i^2(1 - r_{12}^2)(1 - r_{13}^2) = \sigma_i^2(1 - r_{13}^2)(1 - r_{12}^3)
\]

Multiple Correlation coefficient. Let \( R_{1(23)} \) denote the multiple correlation of the variable \( x_1 \) on \( x_2 \) and \( x_3 \). Then it is defined as the correlation between \( x_1 \) and its estimate

\[
X_1 = b_{123}x_2 + b_{132}x_3 = x_1 - x_{1.23}.
\]

Then

\[
R_{1(23)} = \frac{\sum x_iX_i}{\sqrt{\sum x_i^2 \sum X_i^2}} = \frac{\sum x_i(b_{123}x_2 + b_{132}x_3)}{\sqrt{\sum x_i^2 \sum (x_i - x_{1.23})^2}}
\]

\[
= \frac{\sum x_i^2 - \sum x_{1.23}^2}{\sqrt{\sum x_i^2 (\sum x_i^2 - 2\sum x_i x_{1.23} + \sum x_{1.23}^2)}} = \frac{\sum x_i^2 - \sum x_{1.23}^2}{\sqrt{\sum x_i^2 (\sum x_i^2 - \sum x_{1.23}^2)}}
\]

\[
= \frac{N\sigma_i^2 - N\sigma_{1.23}^2}{\sqrt{N\sigma_i^2 (N\sigma_i^2 - N\sigma_{1.23}^2)}} = \frac{\sigma_i^2 - \sigma_{1.23}^2}{\sqrt{\sigma_i^2 (\sigma_i^2 - \sigma_{1.23}^2)}} = \left(1 - \frac{\sigma_{1.23}^2}{\sigma_i^2}\right)^{1/2}
\]

This is the required coefficient of multiple correlation. This result may be expressed as

\[
1 - R_{1(23)}^2 = \frac{\sigma_{1.23}^2}{\sigma_i^2}.
\]

Another notation of \( R_{1(23)} \) are \( R_{1.23} \) and note that \( R_{1(23)} > 0 \) since \( \sigma_{1.23}^2 < \sigma_i^2 \) and if \( R_{1(23)} = 1 \) then \( \sigma_{1.23}^2 = 0 \).

Partial Correlation Coefficient. The partial correlation coefficient of \( x_1 \) and \( x_2 \), keeping \( x_3 \) constant is denoted by \( r_{12.3} \) is the square root of the product of \( b_{12.3} \) and \( b_{21.3} \) i.e.,

\[
r_{12.3} = \sqrt{b_{12.3}b_{21.3}} = \sqrt{\frac{r_{12}r_{13}r_{23}}{1 - r_{23}^2} \frac{r_{21} - r_{23}r_{13}}{1 - r_{13}^2}} = 1 - r_{23}^2 - r_{13}^2 - r_{12}^2 + 2r_{12}r_{13}r_{23}.
\]
4.5.4. Illustrative Examples:

Example 1. From the following table, find

(a) the least square regression equation of $x_1$ on $x_2$ and $x_3$.
(b) Determine $x_1$ when $x_2 = 54$ and $x_3 = 9$.
(c) 

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_2^2$</th>
<th>$x_3^2$</th>
<th>$x_1x_2$</th>
<th>$x_1x_3$</th>
<th>$x_2x_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>57</td>
<td>8</td>
<td>3249</td>
<td>64</td>
<td>3648</td>
<td>512</td>
<td>456</td>
</tr>
<tr>
<td>71</td>
<td>59</td>
<td>10</td>
<td>3481</td>
<td>100</td>
<td>4189</td>
<td>710</td>
<td>590</td>
</tr>
<tr>
<td>53</td>
<td>49</td>
<td>6</td>
<td>2401</td>
<td>36</td>
<td>2597</td>
<td>318</td>
<td>294</td>
</tr>
<tr>
<td>67</td>
<td>62</td>
<td>11</td>
<td>3844</td>
<td>121</td>
<td>4154</td>
<td>737</td>
<td>682</td>
</tr>
<tr>
<td>55</td>
<td>51</td>
<td>8</td>
<td>2601</td>
<td>64</td>
<td>2805</td>
<td>440</td>
<td>408</td>
</tr>
<tr>
<td>58</td>
<td>50</td>
<td>7</td>
<td>2500</td>
<td>49</td>
<td>2900</td>
<td>406</td>
<td>350</td>
</tr>
<tr>
<td>77</td>
<td>55</td>
<td>10</td>
<td>3025</td>
<td>100</td>
<td>4235</td>
<td>770</td>
<td>550</td>
</tr>
<tr>
<td>57</td>
<td>48</td>
<td>9</td>
<td>2304</td>
<td>81</td>
<td>2736</td>
<td>513</td>
<td>432</td>
</tr>
<tr>
<td>56</td>
<td>52</td>
<td>10</td>
<td>2704</td>
<td>100</td>
<td>2912</td>
<td>560</td>
<td>520</td>
</tr>
<tr>
<td>51</td>
<td>42</td>
<td>6</td>
<td>1764</td>
<td>36</td>
<td>2142</td>
<td>306</td>
<td>252</td>
</tr>
<tr>
<td>76</td>
<td>61</td>
<td>12</td>
<td>3721</td>
<td>144</td>
<td>4638</td>
<td>912</td>
<td>732</td>
</tr>
<tr>
<td>68</td>
<td>57</td>
<td>9</td>
<td>3249</td>
<td>81</td>
<td>3876</td>
<td>612</td>
<td>513</td>
</tr>
<tr>
<td>753</td>
<td>643</td>
<td>106</td>
<td>38,843</td>
<td>976</td>
<td>40,830</td>
<td>6,796</td>
<td>5,779</td>
</tr>
</tbody>
</table>


Solution. (a) The linear regression equation of $x_1$ on $x_2$ and $x_3$ is

\[ x_1 = b_{1.23} + b_{12.3}x_2 + b_{13.2}x_3 \]

and normal equations are

\[ \Sigma x_1 = n b_{1.23} + b_{12.3} \Sigma x_2 + b_{13.2} \Sigma x_3, \]
\[ \Sigma x_1 x_2 = b_{1.23} \Sigma x_2 + b_{12.3} \Sigma x_2^2 + b_{13.2} \Sigma x_2 x_3 \]
\[ \Sigma x_1 x_3 = b_{1.23} \Sigma x_3 + b_{12.3} \Sigma x_2 x_3 + b_{13.2} \Sigma x_3^2. \]

Then the normal equations are (here $n = 12$)
12 b_{1,23} + 643 b_{12,3} + 106 b_{13,2} = 753, \ 643 b_{1,23} + 34843 b_{12,3} + 5779 b_{13,2} = 40830 \ and \ 106 b_{1,23} + 5779 b_{12,3} + 976 b_{13,2} = 6796. \ Solving \ these \ equations, \ we \ get \\
\ b_{1,23} = 3.6512, \ b_{12,3} = 0.8546, \ b_{13,2} = 1.5063, \ hence \ x_1 = 3.6512 + 0.8546 x_2 + 1.5063 x_3. \\
(b) \ Substituting \ x_2 = 54 \ and \ x_3 = 9 \ in \ the \ equation \ of \ regression, \ we \ obtain \ x_1 = 63.356. \\

Example 2. Prove that \ \( R_{i(23)}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2} \). \\

Solution. We know that \ \( R_{i(23)}^2 = \left(1 - \frac{\sigma_{123}^2}{\sigma_i^2}\right) \), where \\
\ \( \sigma_{123}^2 = \frac{\sigma_i^2}{1 - r_{23}^2}(1 - r_{23}^2 - r_{12}^2 - r_{13}^2 + 2r_{12}r_{13}r_{23}) \), therefore \\
\ \( R_{i(23)}^2 = 1 - \frac{\sigma_i^2}{1 - r_{23}^2}(1 - r_{23}^2 - r_{12}^2 - r_{13}^2 + 2r_{12}r_{13}r_{23}) = \frac{(1 - r_{23}^2 - r_{12}^2 - r_{13}^2 + 2r_{12}r_{13}r_{23})}{1 - r_{23}^2} \). \\
\ \( \Rightarrow R_{i(23)}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2} \). \\

Example 3. Prove that \ \( b_{12,3} b_{23,1} b_{31,2} = r_{12,3} \ r_{23,1} \ r_{31,2} \). \\

Solution. From the normal equations \\
\ \( \Sigma x_2 \ x_{1,23} = \Sigma x_2, x_{1,23} = \Sigma x_2, (x_1 - \Sigma x_2, x_2 - \Sigma x_2, x_3) = \Sigma x_1 x_{2,3} - b_{12,3} \Sigma x_2 x_{2,3} \) (since \( \Sigma x_2, x_3 = 0 \) \\
\ \( = \Sigma x_1 x_{2,3} - \Sigma b_{12,3} x_{2,3}^2 = 0 \), which gives \( b_{1,23} = \frac{\sum x_{1,23} x_{2,3}^2}{\sum x_{2,3}^2} = \frac{r_{123} \sigma_{13}}{\sigma_{2,3}} \). Similarly \\
\ \( b_{23,1} = \frac{r_{231} \sigma_{2,1}}{\sigma_{3,1}}, b_{31,2} = \frac{r_{312} \sigma_{3,2}}{\sigma_{1,2}} \). Therefore \\
\ \( b_{12,3} b_{23,1} b_{31,2} = \frac{r_{123} \sigma_{1,3}}{\sigma_{2,3}} \cdot \frac{r_{231} \sigma_{2,1}}{\sigma_{3,1}} \cdot \frac{r_{312} \sigma_{3,2}}{\sigma_{1,2}} = r_{12,3} \ r_{23,1} \ r_{31,2} \). \\

Example 4. For a trivariate distribution, show that \ \( 1 - R_{i(23)}^2 = (1 - r_{12}^2)(1 - r_{13,2}^2) \). \\

Solution. We know that \ \( 1 - R_{i(23)}^2 = \sigma_{1,23}^2 / \sigma_i^2 \), where \( \sigma_{1,23}^2 = \sigma_i^2 (1 - r_{12}^2)(1 - r_{13,2}^2) \). Hence \\
\ \( 1 - R_{i(23)}^2 = (1 - r_{12}^2)(1 - r_{13,2}^2) \).
Example 5. If \( r_{12} = 0.86, r_{13} = 0.65, r_{23} = 0.72 \), then find \( r_{12.3} \).

Solution. Since

\[
\begin{align*}
    r_{12.3} &= \sqrt{\frac{r_{12} - r_{13} r_{23}}{1 - r_{23}^2} \times \frac{r_{21} - r_{23} r_{13}}{1 - r_{13}^2}} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1 - r_{23}^2)(1 - r_{13}^2)}} = \frac{0.86 - (0.65)(0.72)}{\sqrt{(1 - (0.65)^2)(1 - (0.72)^2)}} = 0.744 .
\end{align*}
\]

Example 6. If \( r_{12.3} = 0 \) then show that \( r_{13.2} = r_{13} \sqrt{\frac{1 - r_{23}^2}{1 - r_{12}^2}} \).

Solution. Since \( r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1 - r_{23}^2)(1 - r_{13}^2)}} = 0 \Rightarrow r_{12} = r_{13} r_{23} \), therefore

\[
\begin{align*}
    r_{13.2} &= \frac{r_{13} - r_{12} r_{23}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}} = \frac{r_{13} - (r_{13} r_{23}) r_{23}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}} = \frac{r_{13}(1 - r_{23}^2)}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}} = r_{13} \sqrt{\frac{1 - r_{23}^2}{1 - r_{12}^2}} .
\end{align*}
\]

4.6. Summary of Unit:

1. Curve Fitting means an expression of the relationship between two variables by algebraic equations on the basis of observed data.

2. Whenever two variables \( x \) and \( y \) are so related that a change in one is accompanied by change in the other in such a way that an increase in the one is accompanied by an increase or decrease in the other, then variables are said to be correlated. When an increase (or decrease) in one variate corresponds to an increase (or decrease) in the other, the correlation is said to be positive. It is negative when increase in one corresponds to decrease in the other or vice versa.

3. Karl Pearson’s Coefficients of Correlation: The coefficients of correlation \( r \) between two variables \( X \) and \( Y \) is defined by the relation

\[
r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} = \frac{p}{\sigma_x \sigma_y} .
\]

Where \( x \) and \( y \) are the deviations measured from their respective means and \( \sigma_x, \sigma_y \) being the standard deviations of these series.

4. Correlation coefficient and rank correlation coefficients lies between \(-1\) to \(+1\).

5. By regression we mean that an association or relation between two variates \( x \), and \( y \). If \( r = +1 \) or \(-1\), the two regression lines will coincide. The variables are perfectly correlated. If \( r = -1 \), the variables are perfectly negatively correlated, low values of one corresponding to
high values of the other. If \( r = +1 \), variables are perfectly positively correlated, high values of one corresponding to high values of the other. If \( r = 0 \), the two lines of regression become \( X = M_x \) and \( Y = M_Y \), which are two lines parallel to \( Y \) and \( X \) axes respectively, passing through their means \( M_x \) and \( M_y \). They are perpendicular to each other. It means that mean values of \( X \) and \( Y \) do not change with \( y \) and \( X \) respectively \( i.e., X \) and \( Y \) are independent.

(6) The covariance of two independent variates is equal to zero.

(7) (Schwarz’s Inequality): For any stochastic variates \( x \) and \( y \)

\[
[E(xy)]^2 \leq E(x^2)E(y^2).
\]

(8) Sign of correlation coefficient and regression coefficients are same.

(9) The sum of the products of corresponding values of a variate and a residual is zero provided the subscript of the variate occurs among the secondary subscripts of the residual.

(10) The sum of the products of two residuals is unaltered by omitting from the residual any or all of the secondary subscripts which are common to both.

(11) The sum of the products of two residuals is zero provided all the subscripts of one residual occur among the secondary subscripts of the second.

4.7. Assignments

1. Form normal equations and hence find the most plausible values of \( x \) and \( y \) from the following \( x + y = 3.01, 2x - y = 0.03, x + 3y = 7.03, 3x + y = 4.97 \). Ans \( x = 0.999, y = 2.004 \).

2. Fit a second degree parabola to the following data, \( x \) is the independent variable.

<table>
<thead>
<tr>
<th>X</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>2</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>10</td>
<td>11</td>
<td>11</td>
<td>10</td>
<td>9</td>
</tr>
</tbody>
</table>

Ans. \( y = -1 + 3.55x - 0.27x^2 \).

3. Fit a parabolic curve of regression of \( y \) on \( x \) to the seven pairs of values
4. Fit a second degree parabola to the following data:

<table>
<thead>
<tr>
<th>x</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>1090</td>
<td>1220</td>
<td>1390</td>
<td>1625</td>
<td>1915</td>
</tr>
</tbody>
</table>

Ans. $y = 1024 + 40.5x + 27.5x^2$.

5. Derive the least square equations for fitting a curve of the type $y = ax^2 + (b/x)$ to a set of $n$ points. Hence fit a curve of this type to the data:

<table>
<thead>
<tr>
<th>X</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>-1.51</td>
<td>0.99</td>
<td>3.88</td>
<td>7.66</td>
</tr>
</tbody>
</table>

Ans. $y = 0.509x^2 - (2.04/x)$

6. Calculate the value of Pearson's coefficient of correlation for the following series A and B:

<table>
<thead>
<tr>
<th>A</th>
<th>105</th>
<th>104</th>
<th>102</th>
<th>101</th>
<th>100</th>
<th>99</th>
<th>98</th>
<th>96</th>
<th>93</th>
<th>92</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>101</td>
<td>103</td>
<td>100</td>
<td>98</td>
<td>95</td>
<td>96</td>
<td>104</td>
<td>92</td>
<td>97</td>
<td>94</td>
</tr>
</tbody>
</table>

Ans. 0.6

7. Calculate the value of $r$ between X and Y for the values given below:

<table>
<thead>
<tr>
<th>A</th>
<th>1</th>
<th>5</th>
<th>7</th>
<th>9</th>
<th>19</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>25</td>
<td>27</td>
<td>26</td>
<td>29</td>
<td>34</td>
<td>35</td>
</tr>
</tbody>
</table>

Ans. 0.967
8. Calculate the correlation coefficient from the following data:

\( n = 10, \Sigma x = -71, \Sigma x^2 = 555, \Sigma y = 70, \Sigma y^2 = 526, \Sigma xy = 527. \)

9. Calculate the coefficient of correlation for the following table:

<table>
<thead>
<tr>
<th>y( \downarrow ) / x( \rightarrow )</th>
<th>0 - 4</th>
<th>4 - 8</th>
<th>8 - 12</th>
<th>12 - 16</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 5</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 - 10</td>
<td>6</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 - 15</td>
<td></td>
<td>5</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>15 - 20</td>
<td></td>
<td>7</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>20 - 25</td>
<td></td>
<td></td>
<td></td>
<td>9</td>
</tr>
</tbody>
</table>

Ans. 0.8666

10. Find the coefficient of correlation for the following table:

<table>
<thead>
<tr>
<th>x( \rightarrow )</th>
<th>94.5</th>
<th>96.5</th>
<th>98.5</th>
<th>100.5</th>
<th>102.5</th>
<th>104.5</th>
<th>106.5</th>
<th>108.5</th>
<th>110.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>y( \downarrow )</td>
<td>29.5</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>59.5</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>18</td>
<td>6</td>
<td>9</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>89.5</td>
<td>7</td>
<td>3</td>
<td>16</td>
<td>16</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>119.5</td>
<td>5</td>
<td>9</td>
<td>10</td>
<td>9</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>149.5</td>
<td>3</td>
<td>5</td>
<td>8</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>179.5</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>209.5</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>239.5</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Ans. -0.49

11. Calculate the coefficient of correlation from the data given below by the method of differences:
12. Two judges in beauty contest rank the ten competitors as follows:

<table>
<thead>
<tr>
<th>A</th>
<th>6</th>
<th>4</th>
<th>3</th>
<th>1</th>
<th>2</th>
<th>7</th>
<th>9</th>
<th>8</th>
<th>10</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>7</td>
<td>5</td>
<td>8</td>
<td>10</td>
<td>9</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

What is the coefficient of rank correlation? Ans. 0.224

13. Illustrate the methodology of computing the correlation coefficient and the equation of the line of regressions by using the following data:

<table>
<thead>
<tr>
<th>x</th>
<th>2</th>
<th>6</th>
<th>4</th>
<th>7</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>8</td>
<td>8</td>
<td>5</td>
<td>6</td>
<td>2</td>
</tr>
</tbody>
</table>

14. Calculate the regression equation from the following data:

<table>
<thead>
<tr>
<th>Age of Husband</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
<th>26</th>
<th>27</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age of wife</td>
<td>17</td>
<td>17</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>19</td>
<td>19</td>
<td>20</td>
<td>21</td>
<td>22</td>
</tr>
</tbody>
</table>

Ans. $y - 18.9 = 1.42(x - 22.5)$, $x - 22.5 = 1.74(y - 18.9)$

15. The following data are given for marks in English and Mathematics of a certain examination: Mean marks in English = 39.5, Mean marks in Maths.= 47.6, $r = 0.42$, Standard deviation of marks in English = 10.8, Standard deviation of marks in Maths.= 16.9. Form the two lines of regression and calculate the expected average marks in Maths of candidates who received 50 marks in English.

Ans. $y = 0.657x + 21.64$, $y - 47.6 = 0.657(x - 39.5), y = 0.268x + 26.73$, $y - 39.5 = 0.268(y - 47.60); y = 54.5$
16. Two lines of regression are given by \( x + 2y - 5 = 0 \), \( 2x + 3y - 8 = 0 \) and \( \sigma_x^2 = 12 \). Calculate the mean values of \( x \) and \( y \), variance of \( y \) and the coefficient of correlation between \( x \) and \( y \). 

Ans. 1, 2, 0.86, 4

17. The following regression equations have been obtained from a correlation table:

\( y = 0.516x + 33.73 \), \( x = 0.512y + 32.52 \). Find the mean of \( x \)'s and also of \( y \)'s as well as the correlation coefficient between \( x \) and \( y \). What is the ratio of the standard deviation of \( x \) to that of \( y \).

18. For bivariate data: \( n = 18 \), \( \Sigma x^2 = 60 \), \( \Sigma y^2 = 96 \), \( \Sigma x = 12 \), \( \Sigma y = 18 \), \( \Sigma xy = 48 \). Find the equation of regression lines. Ans. \( y = -0.18x - 1.12 \), \( x = -0.12y + 0.89 \).

19. If \( x \) and \( y \) are two correlated variances with the same standard deviation and the coefficient of correlation \( r \), show that the correlation coefficient between \( x \) and \( x + y \) is \( \sqrt{\frac{1}{2}(1 + r)} \).

20. The variates \( x \) and \( y \) have variances \( \sigma_1^2 \) and \( \sigma_2^2 \) respectively and are correlated with coefficient of correlation \( \rho \), \( \xi \) and \( \eta \) defined by \( \xi = x \cos \theta + y \sin \theta \) and \( \eta = y \cos \theta - x \sin \theta \). Show that \( \xi \) and \( \eta \) will be uncorrelated if \( \tan 2\theta = \frac{2\rho \sigma_1 \sigma_2}{\sigma_1^2 - \sigma_2^2} \).

21. If \( x \), \( y \), \( z \) be independent variates each having the same standard deviation. If \( u = x + y \) and \( v = y + z \), prove that the coefficient of correlation between \( u \) and \( v \) is \( \frac{1}{2} \).

22. For any two variates \( x \) and \( y \) \( u = 2x + 7y \) and \( v = 3x + ky \). Then find for what value of \( k \), \( u \) ad \( v \) will be un-correlated.

23. Show that correlation coefficient between the residuals \( x_{1,23} \) and \( x_{2,43} \) is equal and opposite to that between \( x_{1,3} \) and \( x_{2,3} \).

24. In a distribution \( \sigma_1 = 2 \), \( \sigma_2 = \sigma_3 = 3 \), \( r_{12} = 0.7 \), \( r_{23} = 0.5 = r_{31} \), evaluate \( r_{23,1} \), \( R_{1,23} \), \( b_{12,3} \), \( b_{13,2} \), \( \sigma_{1,23} \).

Ans. 0.2425, \( \pm 0.7211 \), 0.4, 0.1333 , 1.3064

24. Show that \( b_{12,3} = \frac{b_{12} - b_{13}b_{23}}{1 - b_{23}b_{32}} \).
25. Find partial correlation coefficients, where \( r_{12} = 0.70, r_{13} = 0.61, r_{23} = 0.4. \)

4.8. Check your progress

1. Find the most plausible values of \( x \) and \( y \) from the four equations:

   (i) \( x + y = 3, x - y = 2, x + 2y = 4, x - 2y = 1. \) Ans. \( x = 2.5, y = 0.7. \)

   (ii) \( x + 2y + z = 1, 2x + y + z = 4, -x + y + 2z = 3, 4x + 2y - 5z = -7. \) Ans. \( x = 1.16, y = -0.76, z = 2.08. \)

   (iii) \( x - y + 2z = 3, 3x + 2y + 5z = 5, 4x + y + 4z = 51, -x + 3y + 3z = 14. \) Ans. \( x = 6.9, y = 3.6, z = 4.14. \)

2. Find the line of fit to the following data:

   \[
   \begin{array}{c|c|c|c|c|c}
   X & 1 & 2 & 3 & 4 & 5 \\
   Y & 4 & 6 & 3 & 5 & 7 \\
   \end{array}
   \]

   Ans. \( y = 3.5 + 0.5x. \)

3. Find the line of fit to the following data:

   \[
   \begin{array}{c|c|c|c|c|c}
   X & 1 & 2 & 3 & 4 & 5 \\
   Y & 5 & 7 & 9 & 10 & 11 \\
   \end{array}
   \]

   Ans. \( y = 3.9 + 1.5x. \)

4. Fit a second degree parabola to the following data:

   \[
   \begin{array}{c|c|c|c|c|c}
   X & 1 & 2 & 3 & 4 & 5 \\
   Y & 25 & 28 & 33 & 39 & 46 \\
   \end{array}
   \]

   Ans. \( y = 22.8 + 1.442x + 0.643x^2. \)

5. Taking 1913 as origin for \( x \)-series fit a straight line to the following data showing the production of a commodity in different years in Punjab:

   \[
   \begin{array}{c|c|c|c|c|c|c}
   Year & x : & 1911 & 1912 & 1913 & 1914 & 1915 \\
   Production & y : & 10 & 12 & 8 & 10 & 14. \\
   \end{array}
   \]

   Ans. \( y = 10.8 + 0.6x. \)

6. The profits £100\( y \), of certain company in the \( x^{th} \) year of its life are given by
Taking \( u = x - 3 \) and \( v = y - 33 \), find the parabolic regression of \( v \) on \( u \) in the form \( v = a + bu + cu^2 \).

Ans. \( y = -0.086 + 5.3u + 0.643u^2 \).

7. The profit of a certain company in the \( x^{th} \) year of its life are given by

\[
\begin{array}{c|c|c|c|c|c}
  x : & 1 & 2 & 3 & 4 & 5 \\
  y : & 25 & 28 & 33 & 39 & 46 \\
\end{array}
\]

Taking \( u = x - 3 \) and \( v = (y - 1650)/50 \). Fit a second degree parabola to these data.

8. Fit a straight line using the method of least squares, and calculate the average rate of growth per week

\[
\begin{array}{c|c|c|c|c|c|c|c|c|c|c}
  \text{Age} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\
  \text{Weight} & 52.5 & 58.7 & 65.0 & 70.2 & 75.4 & 81.1 & 87.2 & 95.5 & 102.2 & 108.4 \\
\end{array}
\]

Ans. \( y = 6.162x + 45.729 \).

9. A manurial experiment on paddy gave the following results:

\[
\begin{array}{c|c|c|c|c}
  \text{Dose of measure (in lbs)} & x : & 0 & 200 & 300 & 600 \\
  \text{Yield per acre (in lbs.)} & y : & 1544 & 1898 & 2133 & 2327 \\
\end{array}
\]

Fit a Parabola of the second degree to represent the data. Ans. \( y = 1547.9 + 378.4x - 40x^2 \)

10. Fit the curve \( y = ae^{bx} \) to the following data, \( e \) being Napierian base, 2.71828:

\[
\begin{array}{c|c|c|c|c|c|c|c|c|c|c}
  x : & 0 & 2 & 4 & \\
  y : & 5.012 & 10 & 31.62 & \\
\end{array}
\]

Ans. \( y = 4.642e^{0.46x} \)

11. Fit the curve \( y = ax^b \) to the following data:

\[
\begin{array}{c|c|c|c|c|c|c|c|c|c|c}
  x : & 1 & 2 & 3 & 4 & 5 & 6 \\
\end{array}
\]
12. Calculate the coefficient of correlation between the values of $X$ and $Y$ given below:

<table>
<thead>
<tr>
<th>$X$</th>
<th>78</th>
<th>89</th>
<th>97</th>
<th>69</th>
<th>59</th>
<th>79</th>
<th>68</th>
<th>61</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>125</td>
<td>137</td>
<td>156</td>
<td>112</td>
<td>107</td>
<td>136</td>
<td>123</td>
<td>108</td>
</tr>
</tbody>
</table>

Ans. 0.957

13. Calculate the coefficient of correlation for the following data:

<table>
<thead>
<tr>
<th>$x$</th>
<th>-10</th>
<th>-5</th>
<th>0</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>5</td>
<td>9</td>
<td>7</td>
<td>11</td>
<td>13</td>
</tr>
</tbody>
</table>

Ans. 0.9

14. Find Karl Pearson's coefficient of correlation from the following index numbers:

<table>
<thead>
<tr>
<th>Wages</th>
<th>100</th>
<th>101</th>
<th>103</th>
<th>102</th>
<th>100</th>
<th>99</th>
<th>97</th>
<th>98</th>
<th>96</th>
<th>95</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost of Living</td>
<td>98</td>
<td>99</td>
<td>99</td>
<td>97</td>
<td>95</td>
<td>92</td>
<td>95</td>
<td>94</td>
<td>90</td>
<td>91</td>
</tr>
</tbody>
</table>

Ans. 0.85

15. Calculate the coefficient of correlation for the following table:

<table>
<thead>
<tr>
<th>$x\downarrow/y\rightarrow$</th>
<th>16 - 18</th>
<th>18 - 20</th>
<th>20 - 22</th>
<th>22 – 24</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 – 20</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>20 – 30</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>30 – 40</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>40 – 50</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>50 – 60</td>
<td></td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>60 – 70</td>
<td></td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Ans. 0.28

16. Calculate the coefficient of correlation from the data given below:
17. Find the correlation coefficient and the equations of regression lines for the following values of \( x \) and \( y \).

\[
\begin{array}{c|c|c|c|c|c|c|c|c|c|c|}
\text{X:} & 76 & 90 & 98 & 69 & 54 & 82 & 67 & 52 \\
\hline
\text{Y:} & 25 & 37 & 56 & 12 & 7 & 36 & 23 & 11 \\
\end{array}
\]

Ans. 0.952

18. You are given the following results for the heights (\( x \)) and weights (\( y \)) of 100 policemen. \( M_x = 68 \) inches, \( M_y = 150 \) lbs., \( \sigma_x = 2.5 \) inches, \( \sigma_y = 20 \) lbs., \( r = 0.6 \). From these data estimate (a) the height of a particular policeman whose weight is 200 lbs. (b) the weight of particular policeman who is 5 feet tall.

Ans. 71.75 inches, 111.6 lbs.

19. Two random variables have the least square regression lines with equations \( 3x + 2y = 26 \) and \( 6x + y = 31 \). Find the mean values and the correlation coefficient between \( x \) and \( y \). Ans. 4, 7, -0.5

20. Given \( x = 4y + 5 \), \( y = kx + 4 \) are the regression lines of \( x \) on \( y \) and \( y \) on \( x \) respectively, show that \( 0 \leq 4k \leq 1 \). If \( k = 1/16 \), find the means of the two variables and the coefficient of correlation between them.

21. Two variates \( x \) and \( y \) have zero means, the same variance \( \sigma^2 \) and zero correlation. Show that \( u = x \cos \alpha + y \sin \alpha \) and \( v = x \cos \alpha - y \sin \alpha \) have the same variance \( \sigma^2 \) and zero correlation.
22. If \( x \) and \( y \) be two independent variates with \( \sigma_x \), \( \sigma_y \) standard deviation respectively, prove that variance of \( ax + by \) is \( a^2\sigma_x^2 + b^2\sigma_y^2 \).

23. If \( \theta \) is the angle between the regression lines of two variates with correlation coefficient \( r \), prove that \( \sin \theta \leq 1 - r^2 \).

24. If \( R_{1,23} = 1 \), prove that \( R_{2,13} = 1 \), \( R_{3,12} = 1 \).

25. Find partial correlation coefficients, where \( r_{12} = 0.80 \), \( r_{13} = -0.40 \), \( r_{23} = -0.56 \).

26. Show that the values \( r_{12} = 0.60 \), \( r_{13} = -0.40 \), \( r_{23} = 0.7 \) are inconsistent.

4.9. Point of Discussion:

At the end of the unit student discuss or seek clarification on some points, if so mention the points:

A: ---------------------------------------------------------------

---------------------------------------------------------------------

---------------------------------------------------------------------

---------------------------------------------------------------------

B: ---------------------------------------------------------------

---------------------------------------------------------------------

---------------------------------------------------------------------

---------------------------------------------------------------------

C: ---------------------------------------------------------------

---------------------------------------------------------------------
4.10. Suggested Study Materials:


UNIT V – SAMPLING THEORY

Structure

5.1 Introduction

5.2 Objectives

5.3 Sampling of Large samples.

5.4 Null and Alternate Hypothesis.

5.5 Errors of First and Second Kinds.

5.6 Level of Significance.
5.1 Introduction

In practical problems the statistician is often confronted with the necessity of discussing universe (or population) of which he can not examine every member. We are thus led naturally to the question. What can be said about a universe of which we can examine only a limited number of sample members? This question is the origin of the theory of sampling. Any collection of individuals under study is said to be population (or universe). The individuals often called the members or the units of the population may be physical objects or measurements expressed numerically or otherwise. A part or small section selected from the population is called sample and the process of such selection is called sampling.

5.2 Objectives
The fundamental object of sampling is to gel as much information as possible of the whole universe by examining only a part of it. An attempt is thus made through sampling to give the maximum information about the parent universe with the minimum effort.

Another object of sampling is to determine the reliability of the estimates when they are obtained. This can be done by drawing successive samples from the same parent universe and comparing the results obtained from different samples.

The population of concrete objects is called an existent population while a hypothetical population may be defined as the collection of all possible ways in which a specified event can happen. The Population of heads and tails obtained by tossing a coin an infinite number of times is a hypothetical one (provided that it does not wear out).

5.3 Sampling of Large samples.

Simple Sampling: By simple sampling we mean random sampling in which each event has the same chance of success and in which the chances of success of each event are independent of the success or failure of events in the preceding trials.

Example 5.3.1: from a pack of cards, the chance to draw a king is \( \frac{4}{54} \);

If we do not replace it, the chance of drawing the king second time is \( \frac{3}{54} \) and so on

This sampling though random is not simple.

However if the card drawn at the first trial was put back in the pack before the next trial, the random sampling would become simple sampling.

Definition 5.3.1: The drawing of samples from a universe whose members possess the attribute A or a (not-A).

Example 5.3.2: From a universe of persons we may have smokers and non-smokers.

The drawing of a member may be called an 'event' or 'trial'.

The possession of a particular attribute A by the individual selected may be called a success.

Mean and Standard Deviation in Simple Sampling of Attributes.

Suppose we take N samples each having \( n \) members.
Let, $p$ be the probability of success of each member and $q$ of failure so that $p + q = 1$. We know that the frequencies of samples with 0, 1, 2…n success are the terms of the binomial expansion of. $N (q + p)^n$ called the sampling distribution of the number of successes in the sample.

The expected value, or mean value of the number of success is $np$.

The variance is $npq$, and

The standard deviation is $\sqrt{npq}$

The mean of the proportion of successes $= np/n = p$ and

The standard deviation of the proportion of successes $= \sqrt{npq}/n = \sqrt{pq}/n$

5.4 Null and Alternate Hypothesis.

Tests of Significance for Large Samples: Suppose a large number of sample is classified according to the frequencies of an attribute,

If the number of successes in a large sample of size $n$ differs from the expected value $np$, by more than $3\sqrt{npq}$ we call the difference highly significant and the truth of the hypothesis is very improbable.

Definition 5.4.1: The testing of a statistical hypothesis is meant a procedure for deciding whether to accept or reject the hypothesis.

Generally we accept the hypothesis as correct and then we calculate $np$, $\sqrt{npq}$, and apply the above test.

Example 5.4.1: A coin is tossed 400 times and it turns up head 216 times. Discuss whether the coin may be unbiased one.

Suppose that the coin is unbiased.

Expected number of heads in 400 tosses $= np = 400 \times \frac{1}{2} = 200$

The deviation of the actual number of head from expected $= 216 - 200 = 16$.

The standard deviation $= \sqrt{npq} = \sqrt{400 \times \frac{1}{2} \times \frac{1}{2}} = 10$. 
The deviation is only 1.6 times the standard deviation and hence it is likely to appear as a result of fluctuations of simple sampling. We conclude that the coin may be taken as unbiased one.

5.5 Errors of First and Second Kinds.

Standard Error: The standard deviation of a sampling distribution of a statistics also called the Standard error.

Therefore frequencies differing from the expected frequency by more than 3 times the standard error are almost certainly not due to fluctuation of samples. It is sometimes written as S. E.

Probable Error: Instead of standard error some authors have used a quantity called probable error which is .67449 times the standard error. It is more easily understood than the standard error by the layman and the business men and universally used in past.

Some Standard Errors

Following are the standard errors of some important parameters when the Parent universe is assumed normal:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Formula</th>
<th>Coefficient of correlation</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$\frac{\sigma}{\sqrt{n}}$</td>
<td></td>
<td>$\frac{1 - r^2}{\sqrt{n}}$</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>$\frac{\sigma}{\sqrt{2n}}$</td>
<td>$\mu_3$</td>
<td>$6\sigma^6$</td>
</tr>
<tr>
<td>Median</td>
<td>$\frac{\sigma}{\sqrt{2n}}$</td>
<td>$\mu_4$</td>
<td>$96\sigma^8$</td>
</tr>
<tr>
<td>Semi inter quartile range</td>
<td>$.7867\sigma$</td>
<td></td>
<td>$\frac{V^2}{2n}$</td>
</tr>
<tr>
<td>Variance</td>
<td>$\frac{\sigma^2}{\sqrt{n}}$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example 5.5.1: Out of a simple sample of 1000 individuals from the inhabitants of a country we find that 36% of them have blue eyes and the remainder has eyes of some other colors. What can we infer about the proportion of blue eyed individuals in the whole population?

Solution: \( p = \text{proportion of individuals having blue eyes} \)

\[
p = \frac{36}{100} = .36 \text{ so that,}
\]

\[
q = 1 - .36 = .64
\]

Assume that the conditions of this problem will give a simple sample.

S.E. of the proportion of individuals having blue eyes

\[
\text{S.E.} = \sqrt{\frac{0.36 \times 0.64}{1000}} = 0.015
\]

Since 3 times the error = .045

Hence taking .36 to be estimate of proportion of families having blue eyes the limit are .36 ± (3 \times 0.015) = .405 and .319 i.e. 40.5% and 31.9%

Example 5.5.2: A sample of 900 days is taken from meteorological records of a certain district and 100 of them are found to be foggy. What are the probable limits to the percentage of foggy days in the district?

Example 5.5.3: A random sample of 500 pineapples was taken from a large consignment and 65 were found to be bad. Show that the S. E. of the proportion of bad ones in a sample of this size is 0.015 and deduce that the percentage of bad pineapples in the consignment almost certainly lies between 8.5 and 17.5.

5.6 Level of significance

Test of Significance in case of attributes:

Procedure for test of significance is as follows:

1- Assume Null hypothesis \( H_0 \) and alternate hypothesis \( H_1 \).

2- Define a test statistic \( Z = \frac{x - np}{\sqrt{npq}} \sim N(0,1) \) and calculate value of \( Z \).

3- Decide the the level of significant.

(i) For \( \alpha = 5\% \) we reject \( H_0 \) if \( |Z| > 1.96 \) two sided.

(ii) For \( \alpha = 1\% \) we reject \( H_0 \) if \( |Z| > 2.58 \) two sided.
(iii) For $\alpha = .27\%$ we reject $H_0$ if $|Z| > 3$ two sided.

Example 5.6.1: A coin is, tossed 1000 times and the head comes out 516 times. Discuss whether the coin is unbiased one?

Suppose that the coin is unbiased $H_0$. $p = 1/2$

Expected number of heads in 1000 tosses $= np = 1000 \times \frac{1}{2} = 500$

The deviation of the actual number of head from expected $= 550 - 500 = 50$

The standard deviation $= \sqrt{npq} = \sqrt{1000 \times \frac{1}{2} \times \frac{1}{2}} = 25$.

The deviation (50) is only 2 times the standard deviation (25) and hence it is likely to appear as a result of fluctuations of simple sampling. We conclude that the coin may be taken as unbiased one.

Example 5.6.2: In some dice throwing experiment, Weldon throw a die 49152 times, and of these 25145, yielded 4, 5 or 6. Is this consistent with the hypothesis that the die was unbiased?

Solution: The probability of throwing 4, 5 or 6 with one die, $p = 3/6 = \frac{1}{2}$.

So $q = 1 - \frac{1}{2} = \frac{1}{2}$.

$n = 49152$, $x = 25145$.

Expected value of number of successes, $E(x) = np = 49152 \times \frac{1}{2} = 24576$

Standard deviation $= \sqrt{npq} = \sqrt{49152 \times \frac{1}{2} \times \frac{1}{2}} = 110.9$

Test statistic $Z = \frac{x - np}{\sqrt{npq}} = \frac{25145 - 24576}{110.9} = \frac{569}{110.9} = 5.1373$

Since $|Z| > 3$, hence hypothesis $H_0$ at .27% rejected. The difference is not due to sampling fluctuations. Hence the data is not consistent with the hypothesis that the die was unbiased.

5.7 Critical region

Confidence Limits for Unknown $P$

Let $x =$ number of successes, $n =$ sample size. $P =$ probability of success or proportion in the population, $Q = 1 - P$, $p = \frac{x}{n} =$ sample proportion. Then
\[ Z = \frac{x - np}{\sqrt{npq}} \sim N(0,1) = \frac{p - P}{\sqrt{PQ/n}} \sim N(0,1) \]

Now, \(|Z| \leq 3\)

\[ \Rightarrow \left| \frac{p - P}{\sqrt{pq/n}} \right| \leq 3 \]

\[ \Rightarrow p - 3 \sqrt{pq/n} \leq P \leq p + 3 \sqrt{pq/n} \]

These are called confidence limits for \(P\) at \(\alpha = 0.27\%\).

However, the limits of \(np\) are \(np - 3 \sqrt{npq} \leq np + 3 \sqrt{npq}\)

Or, \(x - 3 \sqrt{npq} \leq np \leq x + 3 \sqrt{npq}\)

Comparison of two Large Samples

Let two large simple samples of \(n_1\) and \(n_2\) members be taken from two universes. Let these samples give proportion of the attribute A’s as \(p_1\) and \(p_2\) respectively.

We have to find:

Is the difference \(p_1 \sim p_2\) due to fluctuation of simple sampling, the two populations being similar as regards the given attribute A is concerned?

On the hypothesis that the populations are similar us regards the giver, attribute, we can combine that two samples to give an estimate.

If \(p_0\) be this estimate then it is given by

\[ p_0 = \frac{p_1n_1 + p_2n_2}{n_1 + n_2} \]

If \(e_1\) and \(e_2\) be standard errors in the proportion of successes in the two samples then

\[ e_1^2 = \frac{p_0q_0}{n_1}, \quad e_2^2 = \frac{p_0q_0}{n_2} \]

If \(e\) be the standard error in the proportion of successes in the parent universe, it is given by

\[ e^2 = e_1^2 + e_2^2 \]
If $p_1 \sim p_2 < 3e$ it may be due to fluctuations of sampling but if $p_1 \sim p_2 > 3e$, it may be taken a real difference in the population proportions.

Illustrative Examples

Example 5.7.1: In a simple sample of 600 men from a certain large city 400 are found to be smokers. In one of 900 from another large city, 450 are smoker. Do of data indicate that cities are significantly different with respect to prevalence of smoking among men?

Solution Here $n_1 = 600$, $n_2 = 900$, $p_1 = \frac{x_1}{n_1} = \frac{400}{600} = \frac{2}{3}$, $p_2 = \frac{x_2}{n_2} = \frac{450}{900} = \frac{1}{2}$

$$p_0 = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2} \Rightarrow p_0 = \frac{2}{3} \times 600 + \frac{1}{2} \times 900}{600 + 900} = \frac{17}{30}, q_0 = 1 - \frac{17}{30} = \frac{13}{30}$$

$$e^2 = e_1^2 + e_2^2$$

$$= p_0 q_0 \left( \frac{1}{n_1} + \frac{1}{n_2} \right) = \frac{17}{30} \times \frac{13}{30} \left( \frac{1}{600} + \frac{1}{900} \right) = 0.000682 \Rightarrow e = .026$$

Now $p_1 \sim p_2 < 3e$

Therefore it is highly significant. Our hypothesis that the populations are similar is almost certainly wrong.

Example 5.7.2: A machine puts out 16 imperfect articles in a sample of 500. After machine overhauled, puts out 3 imperfect articles in a batch of 100. Has the machine been improved?

Solution- Do as above.

Sampling Distribution of Mean

Suppose that all possible samples of size $n$ are drawn without replacement from a finite population of size $N$. If $\bar{x}$ and $\sigma$ be the mean and s.d. of the sampling distribution and $M$ and $\sigma$ be the mean and s.d. of the population respectively then,

$$\bar{x} = M, \ \sigma = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$
If the population is finite or if sampling is with replacement, we get

\[ \bar{x} = M, \quad \sigma = \frac{\sigma}{\sqrt{n}} \]

Example 5.7.3: population consists of the five numbers 2, 3, 6, 8, 11. We are to take all possible samples of size two which can be drawn from this population with replacement. Find

(a) The mean of population,

(b) The standard deviation of the population,

(c) The mean of the sampling distribution of means,

(d) The standard deviation of the sampling distribution.

Solution:

(a) Mean of population

\[ \bar{x} = M = \frac{2 + 3 + 6 + 8 + 11}{5} = \frac{30}{5} = 6 \]

(b) The standard deviation of the population,

\[ \sigma^2 = \frac{(2 - 6)^2 + (3 - 6)^2 + (6 - 6)^2 + (8 - 6)^2 + (11 - 6)^2}{5} = \frac{54}{5} = 10.8 \]

\[ \sigma = 3.29. \]

(c) The mean of the sampling distribution of means,

The number of samples \( k = 5^2 = 25 \). The following are the samples of size two

\( (2,2), (3,2), (6,2), (8,2), (11,2), \)

\( (2,3), (3,3), (6,3), (8,3), (11,3), \)

\( (2,6), (3,6), (6,6), (8,6), (11,6), \)

\( (2,8), (3,8), (6,8), (8,8), (11,8), \)

\( (2,11), (3,11), (6,11), (8,11), (11,11). \)

Their respective means are

<table>
<thead>
<tr>
<th>2.0</th>
<th>2.5</th>
<th>4.0</th>
<th>5.0</th>
<th>6.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5</td>
<td>3.0</td>
<td>4.5</td>
<td>5.5</td>
<td>7.0</td>
</tr>
</tbody>
</table>
The mean of the sampling distribution is

\[
\frac{\text{Sum of all sample means}}{25} = \frac{150}{25} = 6.0
\]

(d) The variance of the sampling distribution of means is obtained by subtracting the mean 6 from each number in (1), squaring the result, adding all 25 numbers thus obtained and dividing by 25.

\[
\text{Variance of the sampling distribution} = \frac{\text{Sum of all sample means}}{25} = \frac{135}{25} = 5.40
\]

s.d. of the sampling distribution = \(\sqrt{5.40} = 2.32\)

Note: If replacement of numbers is not done, there will be \(\text{ }^1\text{C}_r\) samples i.e.,

(2, 3), (2, 6), (2, 8), (2, 11), (3, 6), (3, 8), (3, 11), (6, 8), (6, 11), (8, 11).

Here (2, 3) and (3, 2) are the same samples etc.

The sample mean of these samples are respectively

\[
2.5 \quad 4.0 \quad 5.0 \quad 6.5 \quad 4.5 \quad 5.5 \quad 7.0 \quad 7.0 \quad 8.5 \quad 9.5
\]

and mean of sampling distribution = \(\frac{2.5 + 4.0 + \ldots + 9.5}{10} = 6.0\)

\[
\text{Variance of sampling distribution} = \frac{(2.5 - 6.0)^2 + (4.0 - 6.0)^2 + \ldots + (9.5 - 6.0)^2}{10} = 4.05
\]

5.8 Tests of Significance Based on \(\chi^2\), \(t\), \(F\) and \(Z\)

Tests of Significance Based on \(\chi^2\)

\(\chi^2\) Statistic The \(\chi^2\) distribution was discovered by Helmert in 1875 and was rediscovered independently by Karl Pearson in 1900.
Definition 5.8.1: Let $x_1, x_2, \ldots, x_n$ be a random sample size $n$ from a normal distribution with mean $\mu$ and variance $\sigma^2$. Then

$$x_i \sim N(\mu, \sigma^2) \text{ for } i = 1, 2, 3, \ldots, n$$

and $z_i = \frac{x_i - \mu}{\sigma} \sim N(0, 1), i = 1, 2, \ldots, n$

and $\sum_{i=1}^{n} z_i^2$ denoted by $\chi^2$ (read as $ki$ square) $\chi^2$ statistic with $(n - 1)$ degrees of freedom.

Application of $\chi^2$ distribution

1. To test significance of sample variance.
2. To test independence of the attributes in contingency table.
3. To compare a number of frequency.
4. To test the goodness of fit.

1. $\chi^2$ Test for testing the significance of sample variance (sample size $n \leq 30$)

Example 5.8.1: The standard deviation of a certain dimension of articles produced by a machine is 7.5 over a long period. A random sample of 25 articles gave a standard deviation of 10.0. Is it justifiable to conclude that the variability has increased? Use 5% level of significance.

Solution: $H_0 = \sigma^2 = 7.5$

$$H_1 = \sigma^2 > 7.5 \text{ (variability has increased)}$$

Test statistic: $\chi^2 = \frac{\sum (x_i - \bar{x})^2}{\sigma^2} \sim \chi^2$ with $(n - 1)$ d.f.

Computation: $\chi^2 = \frac{ns^2}{\sigma^2}$, under $H_0$

$$\chi^2 = \frac{25 \times 10^2}{(7.5)^2} = 44$$

Conclusion: Since the value of $\chi^2$ calculated is greater than tabular value at 5% d.f. which is 36.415 So $H_0$ is rejected. This means there is justification for believing that the variability has increased.
2- Test independence of the attributes in contingency table: **Procedure and terminology for this test is as given below.**

(a) Contingency table – A table consisting t rows and s columns i.e. s x t frequency cells is called contingency table. The grand total N gives the number of individuals.

(b) Expected frequencies – Assuming that a and B are independent then expected frequencies corresponding to cell (I, j) to the observed frequency is given by

\[ e_{ij} = \frac{(A_i)(B_j)}{N} \]

(c) Test-Statistic: \[ \chi^2 = \sum_i \sum_j \frac{O^2_{ij}}{e_{ij}} - N \] with \( (t - 1)(s - 1) \) degree of freedom

(d) Degrees of Freedom: To obtain the number of degrees of freed for \( t \times s \) contingency table we proceed as follows —

Suppose there are \( p \) columns and \( q \) rows. The sum of the cell frequencies in each row is determined as being the border frequency in that row, and similarly for the columns. Hence each of the \( p \) columns and \( q \) rows imposes a constraint. Thus there are \( (p + q) \) constraints. From \( (p + q) \) subtract 1, because they are not algebraically independent, as the sum of the border column equals the sum of the border row i.e., the total frequency. Hence there are \( p + q - 1 \) independent linear constraints. The total number of cells is \( p \times q \).

Hence \[ v = n - k = pq - (p + q - 1) = (p - 1)(q - 1) \]

Linear Constraints: Constraints which involve linear equations \( m \) cell frequencies are called linear constraints.

(e) Conditions for the application of \( \chi^2 \): Following are the condition which should be satisfied before \( \chi^2 \) -test can be applied—

1. In first place \( N \), the total number of frequencies must be Otherwise \( \chi \)'s are not normally distributed. \( N \) should reasonably be at least 50, however few the number of cells.

2. No theoretical cell frequency should be small less than five.

3. The constraints on the cell frequencies, if any should be linear.

4. The members of the sample should be independent.

(f) Conclusion: \( \chi^2 \) (calculated) is less than \( \chi^2 \) table, then hypothesis of independence is accepted.

Illustrative Examples
Example 5.8.2: Show that in a 2x2 contingency table

\[
\begin{array}{cc}
& a & b \\
& c & d \\
\end{array}
\]

\[\chi^2 = \frac{(a + b + c + d)(ad - bc)^2}{(a + b)(c + d)(b + d)(a + c)}\]

Solution: A 2 x 2 contingency table may be arranged as follows—

<table>
<thead>
<tr>
<th>Attribute</th>
<th>B</th>
<th>(\beta)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>a</td>
<td>b</td>
<td>a + b</td>
</tr>
<tr>
<td>(\alpha)</td>
<td>c</td>
<td>d</td>
<td>c + d</td>
</tr>
<tr>
<td>Total</td>
<td>a + c</td>
<td>b + d</td>
<td>a + b + c + d = N</td>
</tr>
</tbody>
</table>

Here observed frequencies are \(a, b, c, d\) and the corresponding expected frequencies under the assumption that the two attributes A and B are independent are as follows

\[
\frac{(a + b)(a + c)}{N}, \quad \frac{(a + b)(b + d)}{N}, \quad \frac{(c + d)(a + c)}{N}, \quad \frac{(c + d)(b + d)}{N}
\]

respectively. Now \(\chi^2\)

\[
\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}
\]

\[
= \sum \frac{(ad - bc)^2}{a + b + c + d \left( \frac{1}{(a + b)(a + c)} \right)}
\]

After solving

\[
= \frac{(ad - bc)^2(a + b + c + d)}{(a + b)(c + d)(b + d)(a + c)}
\]
Example 5.8.3: From the following table, test the hypothesis that the flower colour is independent of flatness of leaf.

<table>
<thead>
<tr>
<th></th>
<th>Flat Leaves</th>
<th>Curled Leaves</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>White Flowers</strong></td>
<td>99</td>
<td>36</td>
<td>135</td>
</tr>
<tr>
<td><strong>Red Flowers</strong></td>
<td>20</td>
<td>5</td>
<td>25</td>
</tr>
</tbody>
</table>

Use the following table giving the values of $\chi^2$ for one degree of freedom for different values of $P$—

<table>
<thead>
<tr>
<th>$P$</th>
<th>5</th>
<th>1</th>
<th>05</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X^2$</td>
<td>.455</td>
<td>2.706</td>
<td>3.841</td>
</tr>
</tbody>
</table>

Solution: On the hypothesis that the flower colour is independent of flatness of leaves, the theoretical frequencies are:

<table>
<thead>
<tr>
<th></th>
<th>Flat Leaves</th>
<th>Curled Leaves</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>White Flowers</strong></td>
<td>100.41</td>
<td>34.59</td>
<td>135</td>
</tr>
<tr>
<td><strong>Red Flowers</strong></td>
<td>18.59</td>
<td>6.41</td>
<td>25</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>119</td>
<td>41</td>
<td>160</td>
</tr>
</tbody>
</table>

$$\chi^2 = (1.4)^2 \left[ \frac{1}{100.41} + \frac{1}{34.59} + \frac{1}{18.59} + \frac{1}{6.41} \right]$$

$$= 1.9881 \{.0099 + .0289 + .0537 + .1560 \}$$
\[ 1.9881 \times 0.2485 = 0.494 \]

The number of degrees of freedom is

\[ (2 - 1) (2 - 1) = 1 \]

The 5% value of \( \chi^2 \) for one degree of freedom is 3.841. The calculated value is much less than this. This comparison leads us to the conclusion that there is no cause to suspect the hypothesis that the flower color is independent of the flatness of the leaf.

Example 5.8.4: The following table shows the result of inoculation against cholera.

<table>
<thead>
<tr>
<th></th>
<th>Non attacked</th>
<th>attacked</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Inoculated</strong></td>
<td>431</td>
<td>5</td>
<td>436</td>
</tr>
<tr>
<td><strong>Non Inoculated</strong></td>
<td>291</td>
<td>9</td>
<td>300</td>
</tr>
</tbody>
</table>

Is there any significant association between inoculation and attack? Given that,

\[ v = 1, P = 0.047 \text{ for } \chi^2 = 3.2 \text{ and } P = 0.069 \text{ for } \chi^2 = 3.3 \]

Solution: Let us find theoretical frequencies and arrange in the following

<table>
<thead>
<tr>
<th></th>
<th>Non attacked</th>
<th>attacked</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Inoculated</strong></td>
<td>431(427.7)</td>
<td>5(8.5)</td>
<td>436</td>
</tr>
<tr>
<td><strong>Non Inoculated</strong></td>
<td>291(294.3)</td>
<td>9(5.7)</td>
<td>300</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>722</td>
<td>14</td>
<td>736</td>
</tr>
</tbody>
</table>

The theoretical frequencies have been shown in brackets. The first theoretical frequency has been obtained
\[
\frac{431}{736} \times 722 = 427.7 \quad \text{Similarly other can be obtain.}
\]

\[
\chi^2 = (3.3)^2 [\frac{1}{427.7} + \frac{1}{8.3} + \frac{1}{294.3} + \frac{1}{5.7}] = 3.28
\]

No. of degrees of freedom = \((2 - 1)(2 - 1) = 1\).

Since \(\chi^2 = 3.2\) corresponds, to \(P = .047\)

\(\chi^2 = 3.3\)corresponds, to \(P = .069\)

Hence by interpolation \(\chi^2 = 3.28\) corresponds, to \(P = .0706\) approximately

Thus if the hypothesis is true, our data gives results which would be obtained about 7 times in hundred trials. We may be unjustified in rejection the hypothesis. So we have to believe that inoculation and attack are associated.

Example 5.8.5: The following table shows the result of hybrids of Dhatura

<table>
<thead>
<tr>
<th></th>
<th>Flowers Violet</th>
<th>Flowers white</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fruits prickly</td>
<td>47</td>
<td>21</td>
<td>68</td>
</tr>
<tr>
<td>Fruits smooth</td>
<td>12</td>
<td>3</td>
<td>15</td>
</tr>
</tbody>
</table>

Using chi-square test, find the association between colour of flowers and character of fruits, given that,

\(\nu = 1, P = .402\) for \(\chi^2 = .7\) and \(P = .399\) for \(\chi^2 = .71\)

Solution

<table>
<thead>
<tr>
<th></th>
<th>Flowers Violet</th>
<th>Flowers white</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fruits prickly</td>
<td>47(48.34)</td>
<td>21(19.66)</td>
<td>68</td>
</tr>
</tbody>
</table>
Fruits smooth

<table>
<thead>
<tr>
<th></th>
<th>12(10.66)</th>
<th>3(4.34)</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>59</td>
<td>24</td>
<td>83</td>
</tr>
</tbody>
</table>

\[
\frac{47}{83} \times 59 = 48.34 \quad \text{Similarly other can be obtaining.}
\]

\[
\chi^2 = (1.34)^2 \left[ \frac{1}{48.34} + \frac{1}{19.66} + \frac{1}{10.66} + \frac{1}{4.84} \right] = .7105
\]

No. of degrees of freedom \( v = (2 - 1)(2 - 1) = 1 \).

Since \( \chi^2 = .7 \) corresponds, to \( P = .402 \)

\( \chi^2 = .71 \) corresponds, to \( P = .399 \)

Hence

\( P = .39885 \) for \( \chi^2 = .7105 \) by extrapolation.

Since \( P > .05 \), the value is not significant \( \text{Le.} \), the hypothesis is correct.

Hence colour of flowers and character of fruits are not associated and the divergence might have happened on account of fluctuation of sampling.

4- \( \chi^2 \)-Test of Goodness of Fit

Example 5.8.6: Five dice were thrown 192 times and the number of tunes 4,5 or 6 were as follows—

<table>
<thead>
<tr>
<th>No. of dice throwing 4, 5, 6</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f )</td>
<td>6</td>
<td>46</td>
<td>70</td>
<td>48</td>
<td>20</td>
<td>2</td>
</tr>
</tbody>
</table>

Calculate \( \chi^2 \).

Solution: Probability of throwing 4, 5 or 6 is \( \frac{3}{6} = \frac{1}{2} \). Therefore the theoretical frequencies getting 5, 4, 3, 2, 1, 0 successes with 5 dice are respectively the successive terms of

\[
192 \left( \frac{1}{2} + \frac{1}{2} \right)^5
\]

Which are as follows-
6, 30, 60, 60, 30, 6, respectively.

Since for the application of $\chi^2$-test number of frequency should be less than 5, hence

<table>
<thead>
<tr>
<th>No. of dice throwing 4, 5, 6</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1 or 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f$</td>
<td>6</td>
<td>46</td>
<td>70</td>
<td>48</td>
<td>22</td>
</tr>
<tr>
<td>$f_i$</td>
<td>6</td>
<td>30</td>
<td>60</td>
<td>60</td>
<td>36</td>
</tr>
</tbody>
</table>

$\chi^2 = \sum \frac{(f - f_i)^2}{f}$

$\chi^2 = \frac{(6 - 6)^2}{6} + \frac{(46 - 30)^2}{30} + \ldots + \frac{(22 - 36)^2}{36}$

$= 0 + 8.53 + 1.66 + 2.4 + 5.44 = 18.03.$

Example 5.8.7: Five dice were thrown 96 times and the number of tunes 4, 5 or 6 were as follows—

<table>
<thead>
<tr>
<th>No. of dice throwing 4, 5, 6</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f$</td>
<td>7</td>
<td>19</td>
<td>35</td>
<td>24</td>
<td>8</td>
<td>3</td>
</tr>
</tbody>
</table>

Calculate $\chi^2$. Find the probability of getting the result.

Solution: Probability of throwing 4, 5 or 6 is $\frac{3}{6} = \frac{1}{2}$. Therefore the theoretical frequencies getting 5, 4, 3, 2, 1, 0 successes with 5 dice are respectively the successive terms of

$96\left(\frac{1}{2} + \frac{1}{2}\right)^5$

Which are as 3, 15, 30, 30, 15, 3, respectively.

Since for the application of $\chi^2$-test number of frequency should be less than 5, hence on regrouping

<table>
<thead>
<tr>
<th>No. of dice throwing 4, 5, 6</th>
<th>5 or 4</th>
<th>3</th>
<th>2</th>
<th>1 or 0</th>
</tr>
</thead>
</table>
\[ \chi^2 = \sum \frac{(f - f_i)^2}{f} \]

\[ \chi^2 = \frac{(26 - 18)^2}{18} + \frac{(35 - 30)^2}{30} + \ldots + \frac{(11 - 18)^2}{18} \]

\[ = 3.555 + 0.833 + 1.200 + 2.722 = 8.31. \]

To find the probability of this result here d.f. \( v = 4 - 1 = 3 \)

Since \( \chi^2 = 7.815 \) corresponds, to \( P = .05 \)

Since \( \chi^2 = 9.837 \) corresponds, to \( P = .02 \) (from table)

Difference in \( \chi^2 \) is 2.022 and \( P \) is – 0.03

Also difference in \( \chi^2 \) from calculated and table value is 8.31 – 7.815 = 0.495

Since for \( \chi^2 = 2.022 \) difference in \( P \) is – 0.03.

So for \( \chi^2 = 0.495 \) difference in \( P \) is \( \frac{-0.03 \times 0.495}{2.022} = -0.007 \)

Corresponding to \( \chi^2 = 8.31 \) value of \( P = .05 - .007 = 0.043. \)

Required probability = .043 = \( \frac{43}{1000} = \frac{4.3}{100} = \frac{1}{25} \) about 1 in 25

Tests of Significance Based on \( t \):

\( t \)-statistic: The static \( t \) was introduced by W S.Gosset in 1908 who wrote under the name “Student”. That is why it is called student’s \( t \) Later on its distribution was rigorously established by Prof. R.A. Fisher in 1926.

Definition: Let \( x_1, x_2, \ldots, x_n \) be a random sample size \( n \) from a normal distribution with mean \( \mu \) and variance \( \sigma^2 \)

Let \( s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{X})^2 \)
where $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$

Then

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \text{ or } \frac{(x - \mu)\sqrt{n}}{s}$$

Denoted by $t$ (read as t-test) is called t-statistic with $(n - 1)$ degrees of freedom.

When $n < 30$. For large $n$, t-statistic tends to standard normal variant.

Application of t-distribution: The t-distribution is used

1- To test significance of the mean of sample.

2- To test the difference between two means or to compare two samples.

3- Sample coefficient of correlation.

4- Sample coefficient of regression

Procedure to test significance of the mean of sample:

Hypothesis: $H_0 = \mu = \mu_0$. $H_1 = \mu > \mu_0$

Test static: $t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \text{ or } \frac{(x - \mu)\sqrt{n}}{s}$

Conclusion: If $p < .05$, we regard that the value of $t$ is significant. If $p < .01$, we regard it as highly significant. A significant value of $t$ throws doubt on the truth of hypothesis.

Illustrative Examples

Example 5.8.8: A machine which produces mica insulating washers of use in electric devices is set to turn out washers having a thickness of 10 mils (1 mil = 0.001 inch). A sample of 10 washers has an average thickness of 9.52 mils with a standard deviation of 0.60 mil. Find out $t$.

Solution: Here $X = 9.52$, $M = 10$, $s = .60$

Test static: $t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{9.52 - 10\sqrt{10}}{.60} = -\frac{48}{.60} = -4 \times 3.16 = 2.528$
Example 5.8.9: Find the Student's t for following variable values in a sample of eight: -4, -2, -2, 0, 2, 2, 3.3 taking the mean of the universe to be zero.

Solution:

<table>
<thead>
<tr>
<th>Serial no</th>
<th>x</th>
<th>$x - \bar{x}$</th>
<th>$(x - \bar{x})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-4</td>
<td>-4.25</td>
<td>18.0625</td>
</tr>
<tr>
<td>2</td>
<td>-2</td>
<td>-2.25</td>
<td>5.0625</td>
</tr>
<tr>
<td>3</td>
<td>-2</td>
<td>-2.25</td>
<td>5.0625</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>-0.25</td>
<td>0.0625</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>1.75</td>
<td>3.0625</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>1.75</td>
<td>3.0625</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>2.75</td>
<td>7.5625</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>2.75</td>
<td>7.5625</td>
</tr>
<tr>
<td>Total</td>
<td>2</td>
<td></td>
<td>49.5000</td>
</tr>
</tbody>
</table>

$\bar{x} = \text{Mean} = \frac{\sum x}{n} = \frac{2}{8} = .25$

$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} = \sqrt{\frac{49.5}{7}} = \sqrt{7.071428} = 2.659$

Hypothesis $H_0 = \text{The mean of universe } M = 0$, we get

$t = \frac{\bar{x} - M}{s/\sqrt{n}} = \frac{(0.25 - 0)\sqrt{8}}{2.659} = .27$

Example 5.8.10: Ten individuals are chosen at random from a population and their heights are found to be in inches 63, 63, 64, 65, 66, 69, 69, 70, 70, 71 discuss the suggestion that the mean height in the universe is 65 inches given that for 9 degrees of freedom the value of Student's $t$ and 5 percent level of significance is 2.262.

Solution: Test static: $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$ or $\frac{(x - \mu)\sqrt{n}}{s}$
Where, $\bar{x}$ and $s$ are to be calculated from the sample values.

For the calculation table is given on next page

\[
t = \frac{\bar{x} - M}{s} \sqrt{\frac{n}{n-1}} = \frac{(67 - 65)\sqrt{10}}{3.13} = 2.02
\]

<table>
<thead>
<tr>
<th>Serial no</th>
<th>$x$</th>
<th>$x - \bar{x}$</th>
<th>$(x - \bar{x})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>63</td>
<td>-4</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>63</td>
<td>-4</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>64</td>
<td>-3</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>65</td>
<td>-2</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>66</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>69</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>69</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>70</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>70</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>71</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>n = 10</td>
<td>$\sum x = 670$</td>
<td>-</td>
<td>$\sum (x - \bar{x})^2 = 88$</td>
</tr>
</tbody>
</table>

Sample mean, $\bar{x} = \frac{\sum x}{n} = \frac{670}{10} = 67$

Sample standard deviation

\[
s = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} = \sqrt{\frac{88}{9}} = 3.13\text{ inches}
\]

$H_0$: the mean of the universe is 65 inches. lCst. Statistic:

The number of degrees of freedom = $v = 10 - 1 = 9$.

Tabulated value for 9 d.f. at 5% level of significance is 2.262.
Since calculated value of \( t \) is less than tabulated value for 9 d.f. (2.02 < 2.262). This error could have arisen due to fluctuations and we may conclude that the data are consistent with the assumption of mean height in the universe of 65 inches.

Example 5.8.11: The nine items of a sample had the following values

\[ 45, 47, 50, 52, 48, 47, 49, 53, 51. \]

Does the mean of the nine items differ significantly from the assumed population mean of 47.5? Given that

\[ v = 8, \ P = .945 \text{ for } t = 1.8 \text{ and } P = .953 \text{ for } t = 1.9 \]

Solution: It can easily be found with simple calculation that

\[
T = \frac{\bar{x} - \mu}{s} \sqrt{n} = \frac{(49.11 - 47.5) \sqrt{9}}{2.62} = 1.84
\]

Degree of freedom = 9 - 1 = 8. Given

\[ P = .945 \text{ for } t = 1.8 \text{ and } P = .953 \text{ for } t = 1.9 \]

Hence by interpolation \( p = .95 \) for \( t = 1.84 \).

Thus the change of getting value of \( t \) greater than observed is 1 - .95 = .05. The probability of getting \( t \) greater in absolute value is 2 x .05 .10 which is greater than .05. This shows that the value of \( t \) is not significant. Hence the same may be a ransom sample from a normal population of mean 47.5

---

<table>
<thead>
<tr>
<th>Red Roses</th>
<th>Yellow Roses</th>
</tr>
</thead>
<tbody>
<tr>
<td>63, 65, 68, 69, 71 and 72</td>
<td>61, 62, 65, 66, 69, 70, 71, 72, 73</td>
</tr>
</tbody>
</table>

\[ v = 14, \ P = .539 \text{ for } t = .10 \text{ and } P = .527 \text{ for } t = .08 \]
<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>x - (\bar{x})</td>
<td>((x - \bar{x})^2)</td>
<td>y</td>
<td>y - (\bar{y})</td>
<td>((y - \bar{y})^2)</td>
</tr>
<tr>
<td>63</td>
<td>-5</td>
<td>25</td>
<td>61</td>
<td>-6.8</td>
<td>46.24</td>
</tr>
<tr>
<td>65</td>
<td>-3</td>
<td>9</td>
<td>62</td>
<td>-5.8</td>
<td>33.64</td>
</tr>
<tr>
<td>68</td>
<td>0</td>
<td>0</td>
<td>65</td>
<td>-2.8</td>
<td>7.84</td>
</tr>
<tr>
<td>69</td>
<td>1</td>
<td>1</td>
<td>66</td>
<td>-1.8</td>
<td>3.24</td>
</tr>
<tr>
<td>71</td>
<td>3</td>
<td>9</td>
<td>69</td>
<td>1.2</td>
<td>1.44</td>
</tr>
<tr>
<td>72</td>
<td>4</td>
<td>16</td>
<td>69</td>
<td>1.2</td>
<td>1.44</td>
</tr>
<tr>
<td>70</td>
<td>1.2</td>
<td>4.84</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>71</td>
<td>3.2</td>
<td>10.24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>72</td>
<td>4.2</td>
<td>17.64</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>73</td>
<td>5.2</td>
<td>27.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>408</td>
<td>0</td>
<td>60</td>
<td>678</td>
<td>0</td>
<td>153.60</td>
</tr>
</tbody>
</table>

Mean height of red roses is \(\bar{x} = \frac{408}{6} = 68\)

Mean height of yellow roses is \(\bar{y} = \frac{678}{10} = 67.8\)

\[
s^2 = \frac{\left(\sum (x - \bar{x})^2 + \sum (y - \bar{y})^2\right)}{n_1 + n_2 - 2}\]

\[
= \frac{\{60 + 153.6\}}{6 + 10 - 2} = \frac{213.6}{14} = 15.257
\]

\(s = 3.906.\)

Now

\[
t = \frac{\bar{x} - \bar{y}}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{68 - 67.8}{3.906} \sqrt{\frac{60}{3.906}} \sqrt{\frac{60}{16}} = .0994
\]

and degrees of freedom = 10 + 6 - 2 = 14.
\( v = 14, P = .539 \) for \( t = .10 \) and \( P = .527 \) for \( t = .08 \)

Hence for \( t = .0994 \), by interpolation \( P = .538 \).

Therefore Fisher's \( P = 2(1 - .538) = .924 \) which is much greater than .05 I Hence the value of \( t \) is not significant. 'Thus there is nothing to suggest that the universes are unlike as regards height i.e., the suggestion that “red rose”, are on the average taller than yellow roses is wrong.

Example 5.8.13: In a rat feeding experiment, the following results were obtained:

<table>
<thead>
<tr>
<th>Diet</th>
<th>Gain in Weight in gm</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>High protein</strong></td>
<td>13 14 10 11 12 16 10 8 11 12 9 12</td>
</tr>
<tr>
<td><strong>Low Protein</strong></td>
<td>7 11 10 8 10 13 9</td>
</tr>
</tbody>
</table>

Investigate if there is any evidence of superiority of one diet over the other. The value of \( t \) for 17 degrees of freedom at 5% level of significance =2.11.

Solution: Do as above. Hint is given below.

<table>
<thead>
<tr>
<th>High Protein</th>
<th>Low Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x )</td>
<td>( x - \bar{x} )</td>
</tr>
<tr>
<td>13</td>
<td>1.5</td>
</tr>
<tr>
<td>14</td>
<td>2.5</td>
</tr>
<tr>
<td>10</td>
<td>-1.5</td>
</tr>
<tr>
<td>11</td>
<td>-0.5</td>
</tr>
<tr>
<td>12</td>
<td>0.5</td>
</tr>
<tr>
<td>16</td>
<td>4.5</td>
</tr>
<tr>
<td>10</td>
<td>-1.5</td>
</tr>
<tr>
<td>8</td>
<td>-3.5</td>
</tr>
<tr>
<td>11</td>
<td>-0.5</td>
</tr>
<tr>
<td>12</td>
<td>0.5</td>
</tr>
</tbody>
</table>
Mean gain in weight on High Protein = 11.5 Mean gain in weight of Low Protein = $9.5/7$

Substituting these values, we get

$$t = \frac{\bar{x} - \bar{y}}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{25}{2} \cdot \frac{68}{7} \cdot \sqrt{\frac{12 \times 7}{535}} \cdot \sqrt{\frac{11 + 7}{119}}$$

Paired Samples: When the size of two samples is the same, ‘t’ can be obtained using

$$\bar{d} = \frac{\sum d}{n} = \frac{50}{10} = \frac{50}{10} \cdot \frac{55}{10} = \frac{1}{n-1} \sum (d - \bar{d})^2 t = \frac{d - 0}{s_d/\sqrt{n}}$$

Illustrative Examples

Example 5.8.14: Ten school boys were given a test in Mathematics. They were given a month’s further tuition and a second test of equal difficulty was held at the end of it. Do the marks give evidence, that the students have from benefited by the extra coaching?

<table>
<thead>
<tr>
<th>Boys :</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marks in Test I</td>
<td>68</td>
<td>25</td>
<td>58</td>
<td>56</td>
<td>64</td>
<td>55</td>
<td>57</td>
<td>69</td>
<td>34</td>
<td>44</td>
</tr>
<tr>
<td>Marks in Test II</td>
<td>71</td>
<td>39</td>
<td>59</td>
<td>59</td>
<td>57</td>
<td>68</td>
<td>69</td>
<td>76</td>
<td>43</td>
<td>39</td>
</tr>
</tbody>
</table>
Solution:

<table>
<thead>
<tr>
<th>Boys</th>
<th>Test I</th>
<th>Test II</th>
<th>d = y - x</th>
<th>d - \bar{d}</th>
<th>(d - \bar{d})^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>68</td>
<td>71</td>
<td>3</td>
<td>-2</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>25</td>
<td>39</td>
<td>14</td>
<td>9</td>
<td>81</td>
</tr>
<tr>
<td>3</td>
<td>58</td>
<td>59</td>
<td>1</td>
<td>-4</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>56</td>
<td>59</td>
<td>3</td>
<td>-2</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>64</td>
<td>57</td>
<td>-7</td>
<td>-12</td>
<td>144</td>
</tr>
<tr>
<td>6</td>
<td>55</td>
<td>68</td>
<td>13</td>
<td>8</td>
<td>64</td>
</tr>
<tr>
<td>7</td>
<td>57</td>
<td>69</td>
<td>-12</td>
<td>7</td>
<td>49</td>
</tr>
<tr>
<td>8</td>
<td>69</td>
<td>76</td>
<td>7</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>34</td>
<td>43</td>
<td>9</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>10</td>
<td>44</td>
<td>39</td>
<td>-5</td>
<td>-10</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>\sum d = 50</td>
<td></td>
<td></td>
<td>\sum (d - \bar{d})^2 = 482</td>
</tr>
</tbody>
</table>

Here \( n = 10 \).

\[ \bar{d} = \frac{\sum d}{n} = \frac{50}{10} = 5 \]

\[ S_d^2 = \frac{1}{n-1} \sum (d - \bar{d})^2 \]

\[ = 53.56 \]

\[ S_d = (53.56)^{\frac{1}{2}} \]

\[ = 7.3 \]

\( H_0 \): Students have not been benefited by the extra coaching i.e. \( \mu_d = 0 \).

\( H_1 \): \( \mu_d > 0 \) (one sided, upper tail of the t-distribution is the critical region)
Test statistic \( t = \frac{d - 0}{s_d / \sqrt{n}} = \frac{(5 - 0)\sqrt{10}}{7.3} = 2.17 \)

Degrees of freedom = 10 - 1 = 9.

For 9 degrees of freedom \( P (t > 1.833) = .05. \)

Conclusion: Since the calculated value of \( t \) is greater than tabulated value of \( t \), the hypothesis \( H_0 \) is rejected. Consequently the coaching may be beneficial to the school boys.

Remark: If we take \( H_0 : \mu_d = 0. \) and \( H_1 : \mu_d \neq 0, \) i.e., there is difference in the marks of two tests. Then two sided critical region.

\[ \text{Here } t_{\text{cal}} < t_{\text{table}} \]

This gives no evidence against the hypothesis \( H_0. \)

Example 5.8.15: The yield of two 'Type 17' and 'Type 51' of grains in pounds per acre in 6 replications are given below. What comments would you make on the difference in the mean yields? You may assume that if there be 5 degrees of freedom and \( P = 0.2, \) it is approximately 1.476.

<table>
<thead>
<tr>
<th>Replication</th>
<th>Yield in Pounds 'types 17'</th>
<th>Yield in Pounds 'Types 51'</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Replication</td>
<td>Yield in Pounds</td>
<td>Difference</td>
</tr>
<tr>
<td>-------------</td>
<td>-----------------</td>
<td>------------</td>
</tr>
<tr>
<td></td>
<td>Type 17</td>
<td>Type 51</td>
</tr>
<tr>
<td>1</td>
<td>20.50</td>
<td>24.86</td>
</tr>
<tr>
<td>2</td>
<td>24.60</td>
<td>26.39</td>
</tr>
<tr>
<td>3</td>
<td>23.06</td>
<td>28.19</td>
</tr>
<tr>
<td>4</td>
<td>29.98</td>
<td>30.75</td>
</tr>
<tr>
<td>5</td>
<td>30.37</td>
<td>29.97</td>
</tr>
<tr>
<td>6</td>
<td>23.83</td>
<td>22.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$d^*$ = mean of difference of the yields $= \frac{\sum d}{n} = \frac{9.86}{6} = 1.634 \text{ pounds}$

$S_d = \text{S.D. of the difference of the yields} = \sqrt{\frac{\sum (d - d^*)^2}{n-1}} = \sqrt{\frac{36.27}{5}} = 2.69 \text{ pounds}$

let us take the null hypothesis that the difference in types has no effect on yields i.e., the population mean of the difference is zero, then this
\[ t = \frac{d' - 0}{s_d / \sqrt{n}} = \frac{(1.634 - 0) \sqrt{6}}{2.69} = 1.489 \]

The value of \( t \) is less than \( t_{0.05} \) for 5 d.f. and therefore the difference is not significant at 5% level but is significant at 20 % level.

Example 5.8.16. The sleep of 10 patients was measured for the effect of the soporific drugs referred to in the following table as Drug A and Drug B. From the data given below show that there is significant difference between the effects of two drugs, on the assumption that different random samples of patients were used to test the two drugs A and B. You may assume that if there be 9 degrees of freedom and \( P = .05 \), \( t = 2.26 \).

<table>
<thead>
<tr>
<th>Patient</th>
<th>Drug A</th>
<th>Drug B</th>
<th>Patient</th>
<th>Drug A</th>
<th>Drug B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+0.7</td>
<td>1.9</td>
<td>6</td>
<td>3.4</td>
<td>4.4</td>
</tr>
<tr>
<td>2</td>
<td>- 1.6</td>
<td>0.8</td>
<td>7</td>
<td>3.7</td>
<td>5.5</td>
</tr>
<tr>
<td>3</td>
<td>- 0.2</td>
<td>1.1</td>
<td>8</td>
<td>0.8</td>
<td>1.6</td>
</tr>
<tr>
<td>4</td>
<td>- 1.2</td>
<td>0.1</td>
<td>9</td>
<td>0</td>
<td>4.6</td>
</tr>
<tr>
<td>5</td>
<td>- 0.1</td>
<td>- 0.1</td>
<td>10</td>
<td>2.0</td>
<td>3.6</td>
</tr>
</tbody>
</table>

ADDITIONAL HOURS OF SLEEP GAINED BY USE OF SPOROFIC DRUGS

Solution: same as above. Table is given for hint.

<table>
<thead>
<tr>
<th>Patient</th>
<th>Drug A</th>
<th>Drug B</th>
<th>d</th>
<th>( d - \bar{d} )</th>
<th>( (d - \bar{d})^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+0.7</td>
<td>1.9</td>
<td>1.2</td>
<td>- .4</td>
<td>.16</td>
</tr>
<tr>
<td>2</td>
<td>- 1.6</td>
<td>0.8</td>
<td>2.4</td>
<td>.8</td>
<td>.64</td>
</tr>
</tbody>
</table>
Test of Significance of Correlation Coefficient (Small Sample)

Let the joint distribution of $X$ and $Y$ two variables be bivariate normal distribution with mean $\mu_1$, $\mu_2$, standard deviations $\sigma_1$, $\sigma_2$ and correlation coefficient $\rho$. Let a random sample $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be drawn from this bivariate normal population. The $t$-test $H_0: \rho = 0$, the population correlation coefficient is zero, we have

$T$-test statistic, $t = \frac{r \sqrt{n - 2}}{\sqrt{1 - r^2}}$ with $(n - 2)$ degree of freedom

$n = \text{sample size (i.e. number of pairs)},$

$r = \text{sample correlation coefficient}.$

Illustrative Examples

Example 5.8.17. A random sample of size 15 from a bivariate normal populatio gave a correlation coefficient $r = 0.5$. Is this indicating the existence of correlation in the population?

Solution: Test statistic, with $(n - 2)$ degree of freedom

$$
\frac{r \sqrt{n - 2}}{\sqrt{1 - r^2}} = \frac{0.5 \sqrt{15 - 2}}{\sqrt{1 - (-0.5)^2}} = \frac{0.5 \sqrt{13}}{\sqrt{1 - 0.25}} = \frac{1.803}{0.866} = 2.082
$$
The tabulated value of $t_{0.05}$ for 13 degrees of freedom is 2.16 which are more than calculated value of $t$ at 5% level of significance. So sample correlation does not warrant the existence of correlation in the population.

Example 5.8.18: *Find the least value of $r$ the sample of 18 pairs from a bivariate normal population, significant at 5 % level, $r$ being the coefficient of correlation of the sample.*

Snedecore’s F- Distribution

To avoid the logarithmic computations in calculating a value of $z$, G.W. Snedecore defined another statistic $F$ by the relation

$$F = \frac{S_1^2}{S_2^2},$$

where $S_1^2 > S_2^2$

And

$$S_1^2 = \frac{\sum(x - \bar{x})^2}{n_1 - 1}, S_2^2 = \frac{\sum(x - \bar{x})^2}{n_2 - 1}$$

Here $\frac{v_1 F}{v_2} = \frac{(n_1 - 1)S_1^2 / \sigma^2}{(n_2 - 1)S_2^2 / \sigma^2}$ where $v_1 = n_1 - 1, v_2 = n_2 - 1$

The numerator and denominator of the second member are independent of $\chi^2$ variates with $v_1$ and $v_2$ degrees of freedom respectively.

Critical region at 5% level of significance is shown by shaded region.
Illustrative Examples

Example 5.8.19: *Two samples of size 9 and 8 give the sum of squares of deviations from their respective means equal to 160 square and 91 squares respectively. Can they be regarded as drawn from the two normal populations with same variance? Given that $F_{0.05}$ for 8 and 7 d.f. is 3.73*

Solution: *Here*

$$S_1^2 = \frac{\sum (x - \bar{x})^2}{n_1 - 1} = \frac{160}{9-1} = 20$$

$$S_2^2 = \frac{\sum (x - \bar{x})^2}{n_2 - 1} = \frac{91}{7} = 13$$

$$F = \frac{S_1^2}{S_2^2} = \frac{20}{13} = 1.54$$

This calculated value of $F$ is less than $F_{0.05}$ for 8 and 7 degrees of freedom i.e., 3.73. Therefore the calculated value of $F$ is not at all significant. Hence the two samples can be regarded as drawn from two normal populations with the same variance.

Example 5.8.20: *Two independent samples of 8 and 7 items respectively had the following values of the variable:*

<table>
<thead>
<tr>
<th>Sample I</th>
<th>9</th>
<th>11</th>
<th>13</th>
<th>11</th>
<th>15</th>
<th>9</th>
<th>12</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample II</td>
<td>10</td>
<td>12</td>
<td>10</td>
<td>14</td>
<td>9</td>
<td>8</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

*Does the estimate of Population variance differ significantly? Given that for 7 degrees of freedom the value of $F$ at 5% level of significance is 4.20 nearly*

Solution:
<table>
<thead>
<tr>
<th>Sample I</th>
<th>Sample II</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x )</td>
<td>9</td>
</tr>
<tr>
<td>11</td>
<td>121</td>
</tr>
<tr>
<td>13</td>
<td>169</td>
</tr>
<tr>
<td>11</td>
<td>121</td>
</tr>
<tr>
<td>15</td>
<td>225</td>
</tr>
<tr>
<td>9</td>
<td>81</td>
</tr>
<tr>
<td>12</td>
<td>144</td>
</tr>
<tr>
<td>14</td>
<td>196</td>
</tr>
<tr>
<td>94</td>
<td>1138</td>
</tr>
</tbody>
</table>

\[
\bar{x} = \frac{94}{8} = 11.75, \quad \bar{y} = \frac{73}{7} = 10.43
\]

\[
\sum (x - \bar{x})^2 = 2 \sum x^2 - 2\bar{x} \sum x + \sum \bar{x}^2
\]

\[
= 1138 - 2 \times \frac{94}{8} \times 94 + 8 \times \left( \frac{94}{8} \right)^2
\]

\[
= 1138 - \frac{94^2}{8}
\]

\[
= 33.5
\]

Similarly, \( \sum (y - \bar{y})^2 = 23.7 \)

\[
S_1^2 = \frac{\sum (x - \bar{x})^2}{n_1 - 1} = \frac{33.5}{7}
\]

\[
S_2^2 = \frac{\sum (x - \bar{x})^2}{n_2 - 1} = \frac{23.7}{6}
\]
This calculated value is less than the value of $F$ at 5% level of significance. Hence differences are not significant. Therefore the samples may well be drawn from the population with same variance.

Example 5.8.21: Two random samples drawn from two normal populations are:

<table>
<thead>
<tr>
<th>Sample I</th>
<th>20</th>
<th>16</th>
<th>26</th>
<th>27</th>
<th>23</th>
<th>22</th>
<th>18</th>
<th>24</th>
<th>25</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample II</td>
<td>27</td>
<td>33</td>
<td>42</td>
<td>35</td>
<td>32</td>
<td>34</td>
<td>38</td>
<td>28</td>
<td>41</td>
<td>43</td>
</tr>
</tbody>
</table>

Obtain the estimates of the variances of the populations and test whether the two populations having the same variance.

Solution: Do as above

Hint: This calculated value of $F$ (=2.14) is less than $F_{0.05}$ (=3.112) at 11 and 9 degrees of freedom. Hence, the hypothesis $H_0: \sigma_1^2 = \sigma_2^2$ may be accepted. Therefore the samples may be regarded as drawn from the populations which have same variance.

Example 5.8.22: For a random sample of 10 pigs, fed on diet A, the increases in weight in pounds in a certain period were

10, 6, 16, 17, 3, 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.

Show that the estimate of population variance in the two samples was not significantly different (for $v_1 = 11$, $v_2 = 9$, the 5% value of $F = 3.112$)

Solution: Same as above

Hint: Since $F_{\text{Cal}} = 2.14 < F_{0.05} = 3.112$ at (11, 9) $H_0: \sigma_1^2 = \sigma_2^2$ accepted.
The estimates of population variance in the two samples are not significantly different.

Example 5.8.23: Show how you would use Student's t-test and Snedecor's F-test to decide whether the following two samples have been drawn from the same normal population. Which of the two tests would you apply first and why?

<table>
<thead>
<tr>
<th></th>
<th>Size</th>
<th>Mean</th>
<th>Sum of Squares of Deviate from the Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample I</td>
<td>9</td>
<td>68</td>
<td>36</td>
</tr>
<tr>
<td>Sample II</td>
<td>10</td>
<td>69</td>
<td>42</td>
</tr>
</tbody>
</table>

Given that $t_{17} (.05) = 2.11$, $F_{9,7} (.05) = 3.4$.

Hint 1- Since $F_{\text{cal}} = 1.04 < F_{9,7} (.05) = 3.4$. Hence $H_0: \sigma_1^2 = \sigma_2^2$ is accepted.

2- Since $t$ calculated is less than $t$ tabulated hence $H_0: \sigma_1 = \sigma_2$ is accepted.

Thus from (1) and (2) we may conclude that the two samples have been drawn from the same normal populations.

Fisher's z-Distribution

Let $x_1, x_2, \ldots x_n$, and Let $y_1, y_2, \ldots y_n$ be the values of two independent random samples with estimated variances $S_1^2$ and $S_2^2$. Suppose we are required to test significance for the difference between two sample variance To do so Fisher has defined a statistic $z$ by the relation

$$z = \frac{1}{2} \log_e F = \frac{1}{2} \log_e \left( \frac{S_1^2}{S_2^2} \right) = \frac{1}{2} \log_e \left( \frac{S_1}{S_2} \right)$$

If we write $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ then $v_1$ and $v_2$ are the degrees of the estimates $S_1^2$ and $S_2^2$

A. Fisher has shown that if the samples came from the same universe, and that the universe is normal, $z$ is distributed according to the law

$$y = y_0 \cdot \frac{e^{v_1 z}}{\left(v_1 e^{2z} + v_2\right)^{\frac{1}{2}(v_1+v_2)}}$$

This distribution is called Fisher's z-distribution.
Significance Test: The hypothesis to be tested is that $S_1^2$ and $S_2^2$ are the estimates of the same population variance. The divergence of the value of $z$ from 0 is the basis of this test.

For $P_z = 0.05$ a value of $z > z_0$ refuses the hypothesis and we conclude that the sample have been taken from populations with different variances.

If $z < z_0$ we say that our hypothesis is correct.

Illustrative Examples

Example 5.8.24: Two gauge operators are tested for precision in making measurements. One operator completes a set of 26 readings with a standard deviation of 1.34 and the other does 34 readings with a standard deviation of 0.98 what is the level of significance of this difference?

You are given that for $v_1 = 25$ and $v_2 = 33, z_{0.05} = 0.306, z_{0.01} = 0.432$

Solution; $n_1 = 26, SD_1 = 1.34, n_2 = 34, SD_2 = 0.98$

$$S_1^2 = \frac{n_1 SD_1^2}{n_1 - 1} = \frac{26 \times (1.34)^2}{26 - 1} = 1.8674$$

$$S_2^2 = \frac{n_2 SD_2^2}{n_2 - 1} = \frac{34 \times (0.98)^2}{34 - 1} = 0.9895$$

$$F = \frac{S_1^2}{S_2^2} = \frac{1.8674}{0.9895} = 1.8770$$

$$z = \frac{1}{2} \log_e F = \frac{1}{2} \log_e \frac{S_1^2}{S_2^2} = \frac{1}{2} \log_e (1.877) = \frac{1}{2} \log_{10} (1.877) \times \log_e 10 = \frac{1}{2} 0.2735 \times 2.3026 = 0.3149$$
Since $z_{cal} > z_{0.05} (0.305)$ and $z_{cal} > z_{0.01} (0.432)$

Hence the difference between variances is significant at 5% level of significance.

Example 5.8.25: Show how you would use Student's $t$-test and Fisher's $z$-test to use whether the two sets of observations:

<table>
<thead>
<tr>
<th>1st Observation</th>
<th>2nd Observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>$x=(x-23)$</td>
</tr>
<tr>
<td>17</td>
<td>-6</td>
</tr>
<tr>
<td>27</td>
<td>4</td>
</tr>
<tr>
<td>18</td>
<td>-5</td>
</tr>
<tr>
<td>25</td>
<td>2</td>
</tr>
<tr>
<td>27</td>
<td>4</td>
</tr>
<tr>
<td>29</td>
<td>6</td>
</tr>
<tr>
<td>27</td>
<td>4</td>
</tr>
<tr>
<td>23</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>-6</td>
</tr>
</tbody>
</table>

Indicate samples drawn from the same universe,

[The value of $z$ at 5% points for 8 and 7 degrees of freedom is .6575 and the value of $z$ at 1% points for 8 and 7 degrees of freedom is 0.9614]

Solution: Method is same as 5.8.24. Table and hint is as given below.
Hint: Since calculated value of \( z \) is more than table value of \( Z \) at 5% level of significance with 8 and 7 degrees of freedom and hence

\[
\text{H}_0 : \sigma_1^2 = \sigma_2^2 \quad \text{is rejected at 5\%.}
\]

Since the calculated value of \( Z \) is less than table value of \( Z \) at 1% level of significance for 8 and 7 degrees of freedom. \( \text{H}_0 : \sigma_1^2 = \sigma_2^2 \) is accepted at 1 % level. The two population variances are same at 1% level of significance.

So we can apply Student's t statistics to test \( \text{H}_0 : \sigma_1 = \sigma_2 \) at 1 % level of significance.

\section*{5.9 Assignment}

1. The following is the distribution of 106 nine pig-litters according to the numbers of males in the litters-

<table>
<thead>
<tr>
<th>No.of Males</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No.of Litters</td>
<td>6</td>
<td>5</td>
<td>8</td>
<td>22</td>
<td>23</td>
<td>25</td>
<td>12</td>
<td>1</td>
<td>4</td>
<td>106</td>
</tr>
</tbody>
</table>

Fit a Binomial distribution under the hypothesis that the sex ratio is 1:1. Test the goodness of fit. Given that \( \chi^2 \) for 4 degrees of freedom at 5% level of significance=9.488.

2. The following table gives the number of aircraft accident that occurred during the various days of the week. Find whether the accident is uniformly distributed over the week.

<table>
<thead>
<tr>
<th>Days</th>
<th>Mon.</th>
<th>Tue.</th>
<th>Wed.</th>
<th>Thur.</th>
<th>Fri.</th>
<th>Sat.</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of Accident</td>
<td>14</td>
<td>16</td>
<td>12</td>
<td>19</td>
<td>9</td>
<td>14</td>
<td>84</td>
</tr>
</tbody>
</table>

[Given that \( \chi^2 \) for 5 degrees of freedom at 5% level of significance=11.07].

3. Records taken of the number of male and female birth in 800 families having four children are as follows-

<table>
<thead>
<tr>
<th>No.of Male Births</th>
<th>No.of Female Births</th>
<th>No.of Families</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4</td>
<td>32</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>178</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>290</td>
</tr>
</tbody>
</table>
Test whether the data are consistent with the hypothesis that the binomial law holds and that the chance of a male birth is equal to that of a female birth, that is \( q = p = 1/2 \). You may use the table given below-

<table>
<thead>
<tr>
<th>Degrees of Freedom</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>5% Value of ( \chi^2 )</td>
<td>3.84</td>
<td>5.99</td>
<td>7.82</td>
<td>9.49</td>
<td>11.07</td>
</tr>
</tbody>
</table>

4. The following data shows the suicides of women in eight German states during fourteen years-

<table>
<thead>
<tr>
<th>No. of suicides in a state per year</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>ObservedFrequency</td>
<td>9</td>
<td>19</td>
<td>7</td>
<td>20</td>
<td>15</td>
<td>11</td>
<td>8</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>112</td>
</tr>
</tbody>
</table>

Fit a Poisson distribution and test its goodness of fit. The value of \( \chi^2 \) for 6 d.f. at 5% level of significance=12.592

5. Genetic theory states that children having one parent of blood type M and the other of blood type N will always be one of the three types M, MN, N and that the proportions of three types will be on average as 1:2:1. A report states that out of 300 children having one M parent and N parent 30\% were found to be type M, 45\% type MN and remainder type N. Test the hypothesis by \( \chi^2 \) for 2 degrees of freedom at 5\% level 5.991.

6. The following table occurs in a memoir of Karl Pearson:

<table>
<thead>
<tr>
<th>Eye colour in father</th>
<th>Eye colour in sons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not light</td>
</tr>
<tr>
<td>Not light</td>
<td>230</td>
</tr>
<tr>
<td>Light</td>
<td>300</td>
</tr>
</tbody>
</table>

Test whether the colour of the son’s eyes is associated with that of the father’s (\( \chi^2 = 3.84, v = 1 \))
7. In an experiment on the immunization of goats from anthrax the following results were obtained. Derive your inference on the efficiency of the vaccine-

<table>
<thead>
<tr>
<th></th>
<th>Died of Anthrax</th>
<th>Survived</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inoculated with Vaccine</td>
<td>2</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>Not Inoculated</td>
<td>6</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>Total</td>
<td>8</td>
<td>16</td>
<td>24</td>
</tr>
</tbody>
</table>

The value of $\chi^2$ for 1 degree of freedom at 5% is 3.481.

8. The following table gives the series of controlled experiment. Discuss whether the treatment may be considered to have any positive effect-

<table>
<thead>
<tr>
<th></th>
<th>Positive</th>
<th>No effect</th>
<th>Negative</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>9</td>
<td>2</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>Control</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>Total</td>
<td>12</td>
<td>8</td>
<td>4</td>
<td>24</td>
</tr>
</tbody>
</table>

The value of $\chi^2$ for 2 degrees of freedom at 5% level of significance is 5.99.

9. In an experiment on immunization of cattle from tuberculosis, the following results were obtained-

<table>
<thead>
<tr>
<th></th>
<th>Affected</th>
<th>Unaffected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inoculated</td>
<td>12</td>
<td>26</td>
</tr>
<tr>
<td>Not inoculated</td>
<td>16</td>
<td>6</td>
</tr>
</tbody>
</table>

Examine the effect of vaccine in controlled susceptibility to tuberculosis.

10. Find the value of $\chi^2$ for the following table-

<table>
<thead>
<tr>
<th>Diet</th>
<th>Males</th>
<th>Females</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>123</td>
<td>153</td>
<td>276</td>
</tr>
<tr>
<td>B</td>
<td>145</td>
<td>150</td>
<td>295</td>
</tr>
<tr>
<td>Total</td>
<td>268</td>
<td>303</td>
<td>571</td>
</tr>
</tbody>
</table>

11. Five’s computed for four fold tables is independent replications of an experiment are 0.50, 4.10, 1.20, 2.79 and 5.41. Does the aggregate of these tests yield a significant? Given-
12. Write a short note on application of $\chi^2$ in tests of significance.

13. What is the use of chi-square distribution in tests of goodness of fit?

14. Explain briefly the uses and limitations of $\chi^2$–test.

15. Write a short note on tests based on $\chi^2$–distribution.

16. Explain any two applications of $\chi^2$–test.

17. Discuss conditions for applications of $\chi^2$–test.

18. Write down an expression for testing the independence of two attributes.

**ANSWERS 5.9**

1. $f_e$: 0.4,3.3,11.6,23.2,29,23.2,11.6,3.3,0.4; $\chi^2=2.52, v=4$. fit is good

2. Accidents appears to be uniformly distributed over all days of the week.

3. $\chi^2 =19.63 > \chi_{0.05}^2 (4)$.

4. Theoretical frequencies are: 4,12,21,24,21,15,8,4,2,1,0. $\chi^2 =18.35, v=6$.

5. $f_e$: 75,150,75; $\chi^2 =4.5$. Appears to be associated.

6. Survival is not associated with inoculation of vaccine.

9. Vaccine is effective.

5.10 Check your progress

1. Describe how to test the significance of an observed correlation coefficient when the corresponding population value is 0.

2. A random sample of 11 observations from a bivariate population gave a correlation coefficient 0.239. Could the observed value have arisen from an uncorrelated population?

   Ans. $T= 0.74$, $t_{0.05} (9) = 2.262$, $H_0$ is accepted

3. A random sample of size 15 from a bivariate normal population gave a correlation coefficient of – 0.5. Is this an indication of the existence of correlation in the population?

   Ans. $T = -2.082$ is not significant

4. Show that in the random sample of size 25 from an uncorrelated normal population the chance is 1 in 100 that $r$ is greater than about 0.43.
5. Find the least value of \( r \) in a sample of 27 paired observations from a bivariate normal population that is significant at 5\% level of significance.

Ans. \( | r | > 0.38 \)

6. Discuss any two tests of significance based on t-distribution

7. Calculate the value of \( t \) in the case of two characters \( A \) and \( B \) whose corresponding value:
   are given below:
   
<table>
<thead>
<tr>
<th></th>
<th>16</th>
<th>10</th>
<th>8</th>
<th>9</th>
<th>9</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( B )</td>
<td>8</td>
<td>4</td>
<td>5</td>
<td>9</td>
<td>12</td>
<td>4</td>
</tr>
</tbody>
</table>

Ans. \( t = 1.66 \).

8. The figures below are for protein tests of the same variety of wheat grown in two districts. The average in District I is 12.74 and in District II is 13.03. Calculate \( r \) for testing the significance between the means of the two districts:

<table>
<thead>
<tr>
<th>Protein results</th>
</tr>
</thead>
<tbody>
<tr>
<td>District</td>
</tr>
<tr>
<td>I</td>
</tr>
<tr>
<td>II</td>
</tr>
</tbody>
</table>

Ans. \( t = 0.85 \)

9. In a Test Examination given to two groups of students the marks obtained were as follows:

<table>
<thead>
<tr>
<th>First group</th>
<th>18</th>
<th>20</th>
<th>36</th>
<th>50</th>
<th>49</th>
<th>36</th>
<th>34</th>
<th>49</th>
<th>41</th>
</tr>
</thead>
<tbody>
<tr>
<td>Second group</td>
<td>29</td>
<td>28</td>
<td>26</td>
<td>35</td>
<td>30</td>
<td>44</td>
<td>46</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Examine the significance of difference between the arithmetic averages of the marks secured by the students of the above two groups.

(The value of \( t \) for 14 degrees of freedom at 5\% level of significance = 2 .14.

Ans. Not significant

10. For a random sample of 12 boys fed on diet \( A \), the increases in weight in pounds in a certain period were
25, 32, 30, 34, 24, 25, 14, 32, 24, 30, 31, 35.

For another random sample of 15 boys fed on diet B, the increase weight in pounds in the same period were
44, 34, 22, 10, 47, 31, 40, 30, 32, 35, 18, 21, 35, 29, 22.

Find whether diet B is superior to diet A.

Given that the value of $t$ for 25 degrees of freedom at 5% level of significance is 2.06.

Ans. Diet B is superior to diet A

11. The means of two random samples of sizes 9 and 7 respectively, are 196.4 and 198.82 respectively. The sum of the squares of the deviations from the means are 26.94 and 18.73 respectively. Can the samples be considered to have been drawn from the same normal population?

It being given that the value of $t$ for 14 d.f. at 5% level of significance is 2.145 and at 1% level of significance is 2.977.

Ans. $t = 2.65$

12. Two types of batteries-A and B are tested for their length of life and following results are obtained:

<table>
<thead>
<tr>
<th>No of Samples</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>10</td>
<td>500 hours</td>
</tr>
<tr>
<td>$B$</td>
<td>10</td>
<td>560 hours</td>
</tr>
</tbody>
</table>

Is their a significant difference in the two means?

The value of $t$ for 18 degrees of freedom at 5% level of significance is 2.1.

Ans. Not significant

5.11 Point for discussion/clarification
At the end of the unit student discuss or seek clarification on some points, if so mention the points:
5.12 Suggested Study material


*****