CHAPTER-1

Classification, Tabulation and Presentation of Data

The raw data collected through surveys or experiments are generally in haphazard and unsystematic form. Such data are not appropriately formed to draw right conclusions about the group or population under study. Hence, it becomes necessary to arrange or organize data in a form, which is suitable for identifying the numbers of units belonging to a more group, for comparisons and for further statistical treatment or analysis of data etc.

According to L.R. Connor, “classification is the process of arranging things (either actually or notionally) in groups or classes according to their resemblance and affinities, and gives expression to the unity of attributes that may subsist amongst the diversity of individuals”.

Norms for an ideal classification

1. The classes should be complete and non-overlapping.
2. Clarity of classes is another important property. One should use standardized classes so that the comparison of results can be possible from time to time

The classification of data

Geographical classification: – Data are arranged according to places, areas or regions.

Chronological classification: - Data are arranged according to time i.e. weekly, monthly, quarterly, half yearly, annually, quinquennially etc.

Qualitative classification: - The data are arranged according to attributes like sex, marital status, educational standard, stage or intensity of disease etc.

Manifold classification: - The numbers of classes is more than two.

Quantitative classification: - Arranging the data according to certain characteristics that has been measured e.g. according to height, weight, or income of person, vitamin content in substance etc.

A numerical formula as suggested by H.A. Struges may be used for Class interval and number of classes :

\[ i = \frac{L-S}{1+3.3222 \log_{10} n} \]
Where, \( 1 + 3.322 \log_{10} n = k \), the number of classes

\( i \) = class interval

\( L \) = larger observation

\( S \) = small observation

\( n \) = total numbers of observation

**Time series:** - Another type of classification in which data or the derived values from data for each time period is arranged in chronologically.

**Tabulation**

Tabulation follows classification. It is logical or systematic listing of related data in row and columns. The row of a table represents the horizontal arrangement of data and the columns are representing vertical arrangement of data.

**Objective of tabulation**

i. Simplify complex data
ii. Highlights important characteristics
iii. Present data in minimum space
iv. Facilitate comparisons
v. Brings out trends and tendencies
vi. Facilitate further analysis

**Basic difference between classification and tabulation**

**Classification**

- It is basis for tabulation
- It is basis for simplification
- Data is divided into groups and subgroups on the basis of similarities and dissimilarities

**Tabulation**

- It is basis for further analysis
- It is basis for presentation
- Data is listed according to a logical sequence of related characteristics
Table: Percentage of COA employees in age group

<table>
<thead>
<tr>
<th>Departments</th>
<th>Age</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>18-38</td>
<td>38 &amp; above</td>
</tr>
<tr>
<td>Agronomy</td>
<td>1.5</td>
<td>3.7</td>
</tr>
<tr>
<td>Horticulture</td>
<td>2.0</td>
<td>3.9</td>
</tr>
<tr>
<td>Entomology</td>
<td>1.5</td>
<td>4.4</td>
</tr>
<tr>
<td>Pathology</td>
<td>2.0</td>
<td>4.0</td>
</tr>
<tr>
<td>Soil science</td>
<td>1.8</td>
<td>5.2</td>
</tr>
<tr>
<td>Total</td>
<td>8.8</td>
<td>21.2</td>
</tr>
</tbody>
</table>

Source: [Teaching Manual on "Statistical Methods for Applied Sciences" by Ravi R. Saxena, M.L. Lakhera and Roshan Bhardwaj, Department of Agricultural Statistics and SS (L), IGKV, Raipur – 492 012 (Chhattisgarh)]

Parts of tables

i. **Table number:** To identify the table for reference
ii. **Title:** Title indicate the scope and the nature of contents in concise form
iii. **Caption:** Caption are the heading and subheading describing the data present in the columns
iv. **Stubs:** Stubs are the heading and subheading of row
v. **Body of the table:** It contains numerical information
vi. **Ruling and Spacing:** Ruling and Spacing separate columns and rows.
vii. **Head Note:** It is given below the title of the table to indicate the unit of measurement of the data and is enclosed in brackets
viii. **Source Note:** Source note indicates the source from which data is taken. It is related to table is placed at the bottom on the left hand corner

Guidelines for preparing a table

i. The shape and size of the table should contain the required number of row and columns with stubs and captions and the whole data should be accommodated within the cells formed corresponding to the rows and column.
ii. If the quantity is the zero, it should be centered as zero. Leaving blank space or putting dash in place of zero is confusing and undesirable.
iii. In case two or more figures are the same, ditto marks should not be used in a table in the place of the original numerable.

iv. The unit of measurement should either be given in parentheses just below the column’s caption or in parentheses along with the stub in the row.

v. If any figure in the table has to be specified for a particular purpose, it should be marked with an asterisk or a dagger. The specification of the marked figures should be explained at the foot of the table with the same mark.

**Cross classification**

In many situations the data are cross classified with regards to two or more classification or *polytomies*. Double classification is most common because such tabulation is very convenient and informative.

*Example*: Farmer’s earning by fields products and waiting time for the income

<table>
<thead>
<tr>
<th>Fields of product</th>
<th>Initial</th>
<th>After second year’s</th>
<th>After five year’s</th>
<th>Waiting time for first income</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paddy</td>
<td>53</td>
<td>87</td>
<td>124</td>
<td>0.92</td>
</tr>
<tr>
<td>Wheat</td>
<td>47</td>
<td>85</td>
<td>123</td>
<td>1.15</td>
</tr>
<tr>
<td>Gram</td>
<td>45</td>
<td>83</td>
<td>122</td>
<td>1.46</td>
</tr>
<tr>
<td>Soybean</td>
<td>43</td>
<td>60</td>
<td>103</td>
<td>1.63</td>
</tr>
<tr>
<td>Peagon pea</td>
<td>36</td>
<td>63</td>
<td>100</td>
<td>1.17</td>
</tr>
</tbody>
</table>

**Presentation of data**

Though the data presented in the form of table yields good information to the educated people, they are not always good for the lay man as the table cannot be understood easily by them. Diagrammatic and graphical presentations are more intelligible, attractive, and appealing. The diagrammatic representations give a bird’s eye-view of the data.

Generally, the diagrams, charts, or graph are two dimensional, though at times they can be three dimensional. In a two-dimensional figure are there are always two axes named as *X* – axis or *abscissa* and *Y*-axis or *ordinate*.

**Diagrammatic representation of data**

**Line and Bar Diagram**: Such diagrams are suitable for discrete variables i.e. for data given according to some periods, places and timings. These periods, places or timings are
represented on the base line (X-axis) at regular intervals and the corresponding values or frequencies are represented on the Y-axis (ordinates)

**Histogram:** The frequency distribution is represented by a set of rectangular bars with area proportional to class frequency. If the class intervals have equal width then the variable is taken along X-axis and frequency along Y-axis and rectangle is construction.

**Frequency Polygon:** The mid value of class intervals are plotted against frequency of the class interval. These points are joined by straight lines and hence the frequency polygon is obtained.

**Frequency Curve:** First we draw histogram for the given data. Then we join the mid points of the rectangles by a smooth curve. Total area under frequency curve represents total frequency. They are most useful form of frequency distribution.

**Ogives:** Ogives is obtained by drawing the graph of cumulative frequency distribution, hence, ogives are also called cumulative frequency curves. Since a cumulative frequency distribution can be of ‘less than’ or ‘greater than’ type, we have less than and greater than type of ogives.

*Less than ogive* – Variable are taken along X-axis and less than cumulative frequencies are taken along Y-axis. Less than cumulative frequency plotted against upper limit of class interval and joined by smooth curve.

*More than ogive* – more than cumulative frequencies plotted against lower limit of the class interval and joined by a smooth curve. From the meeting point these two ogives if we draw a perpendicular to X-axis, the point where it meets X-axis gives median of the distribution.

**Boxplot Basics**

A box plot splits the data set into quartiles. The body of the box plot consists of a "box" (hence, the name), which goes from the first quartile (Q1) to the third quartile (Q3). Within the box, a vertical line is drawn at the Q2, the median of the data set. Two horizontal lines, called whiskers, extend from the front and back of the box. The front whisker goes from Q1 to the smallest non-outlier in the data set, and the back whisker goes from Q3 to the largest non-outlier.
If the data set includes one or more outliers, they are plotted separately as points on the chart. In the box plot above, two outliers precede the first whisker; and three outliers follow the second whisker.

**How to Interpret a Box plot**

Here is how to read a box plot. The median is indicated by the vertical line that runs down the center of the box. In the box plot above, the median is about 400. Additionally, box plots display two common measures of the variability or spread in a data set.

* **Range.** If you are interested in the spread of all the data, it is represented on a box plot by the horizontal distance between the smallest value and the largest value, including any outliers. In the box plot above, data values range from about -700 (the smallest outlier) to 1700 (the largest outlier), so the range is 2400. If you ignore outliers, the range is illustrated by the distance between the opposite ends of the whiskers - about 1000 in the box plot above.

* **Inter quartile range (IQR).** The middle half of a data set falls within the inter quartile range. In a box plot, the inter quartile range is represented by the width of the box (Q3 minus Q1). In the chart above, the inter quartile range is equal to 600 minus 300 or about 300.

And finally, box plots often provide information about the shape of a data set. The examples below show some common patterns.
Each of the above box plots illustrates a different skewness pattern. If most of the observations are concentrated on the low end of the scale, the distribution is skewed right; and vice versa. If a distribution is symmetric, the observations will be evenly split at the median, as shown above in the middle figure.

Consider the box plot below.

Which of the following statements are true?

I. The distribution is skewed right.
II. The interquartile range is about 8.
III. The median is about 10.

(A) I only (B) II only (C) III only (D) I and III (E) II and III

Solution

The correct answer is (B). Most of the observations are on the high end of the scale, so the distribution is skewed left. The interquartile range is indicated by the length of the box, which is 18 minus 10 or 8. And the median is indicated by the vertical line running through the middle of the box, which is roughly centered over 15. So the median is about 15.
CHAPTER - 2

Descriptive Statistics and Exploratory Data Analysis

The collected data as such are not suitable to draw conclusions about the mass from which it has been taken. Some inferences about the population can be drawn from the frequency distribution of the observed values. This process of condensation of data reduces the bulk of data, and the frequency distribution is categorized by certain constraints known as parameters. Generally, a distribution is categorized by two parameters viz., the location parameter (central values) and the scale parameters (measures of dispersion). Hence in finding a central value, the data are condensed into a single value around which the largest number of values tend to cluster, commonly such a value lies in the centre of distribution and is termed as central tendency.

There are three popular measures of central tendency namely, (i) mean, (ii) median and (iii) mode. Each of these will be discussed in detail here. Beside these, some other measures of location are also dealt with, such as quartiles, deciles and percentiles.

Characteristics: It should possesses the following characteristics

1. It should not be affected by the extreme values
2. It should be as close to the maximum number of observed values as possible
3. It should be defined rigidly which means that it should have no discretion.
4. It should not be subjected to complicated and tedious calculations.
5. It should be capable of further algebraic treatment.
6. It should be stable with regards to sampling.

Mean

There are three types of means which are suitable for a particular type of data. They are Arithmetic mean or Average, Geometric mean and Harmonic mean.

Arithmetic mean (AM):

It is also popularly known as average. If mean is mentioned, it implies arithmetic mean, as the other means are identified by their full names. It is the most commonly used measure of central tendency.

Definition: Sum of the observed values of a set divided by the number of observations in the set is called a mean or an average.

If $X_1, X_2, \ldots, X_N$ are $N$ observed values, the mean or average is given as,

$$\mu = \frac{X_1 + X_2 + \ldots + X_n}{N}$$
for \( i = 1, 2, \ldots N \)

\[
= \frac{1}{N} \sum_i X_i
\]

Population mean is usually denoted by \( \mu \) or \( \bar{X} \) whereas, the sample mean is denoted by \( \bar{x} \) (small letter).

When the data are arranged or given in the form of frequency distribution i.e. there are \( k \) variate values such that a value \( X_i \) has a frequency \( f_i \) (\( i = 1, 2, \ldots, k \)) the formula for the mean is

\[
\mu = \frac{\sum_i f_i X_i}{\sum_i f_i}
\]

for \( i = 1, 2, \ldots k \)

\[
= \frac{1}{N} \sum_i f_i X_i
\]

where, \( N = f_1 + f_2 + \ldots f_k = \sum f_i \).

If the data are given with \( k \) class intervals i.e. the data are in the form as follows:

<table>
<thead>
<tr>
<th>Class interval</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X_1 - X_2 )</td>
<td>( f_1 )</td>
</tr>
<tr>
<td>( X_2 - X_3 )</td>
<td>( f_2 )</td>
</tr>
<tr>
<td>( X_3 - X_4 )</td>
<td>( f_3 )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( X_k - X_{k+1} )</td>
<td>( f_k )</td>
</tr>
</tbody>
</table>

The arithmetic mean

\[
\mu = \frac{\sum_i f_i Y_i}{N}
\]

where \( Y_i \) is the mid point of the class interval \( X_i - X_{i+1} \) and is given by

\[
Y_i = \frac{X_i + X_{i+1}}{2}
\]

for \( i = 1, 2, \ldots k \)
Example: The distribution of tree age at plantation of 130 plants was as given below:

Age in days: 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29

No. of plants: 2, 1, 4, 8, 10, 12, 17, 19, 18, 14, 13, 12

The average age can be computed by the formula

\[ A = \frac{18 \times 2 + 19 \times 1 + \ldots + 29 \times 12}{2 + 1 + \ldots + 12} \]

\[ = \frac{3240}{130} \]

\[ = 24.92 \]

The mean age of plants at plantation time is 24.92 days.

Example:

The weight of eighty insects obtained in a life testing experiment has been presented below in the form of “less than” type of distribution.

<table>
<thead>
<tr>
<th>Life of insects (days)</th>
<th>No of insects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 1</td>
<td>3</td>
</tr>
<tr>
<td>“ “ 2</td>
<td>12</td>
</tr>
<tr>
<td>“ “ 3</td>
<td>14</td>
</tr>
<tr>
<td>“ “ 4</td>
<td>22</td>
</tr>
<tr>
<td>“ “ 5</td>
<td>33</td>
</tr>
<tr>
<td>“ “ 6</td>
<td>46</td>
</tr>
<tr>
<td>“ “ 7</td>
<td>58</td>
</tr>
<tr>
<td>“ “ 8</td>
<td>66</td>
</tr>
<tr>
<td>“ “ 9</td>
<td>75</td>
</tr>
<tr>
<td>“ “ 10</td>
<td>80</td>
</tr>
</tbody>
</table>
The given distribution with regular class intervals and their mid-values can be written as,

<table>
<thead>
<tr>
<th>Days</th>
<th>Mid values ($y$)</th>
<th>No. of insects ($f$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-1</td>
<td>0.5</td>
<td>3</td>
</tr>
<tr>
<td>1-2</td>
<td>1.5</td>
<td>9</td>
</tr>
<tr>
<td>2-3</td>
<td>2.5</td>
<td>2</td>
</tr>
<tr>
<td>3-4</td>
<td>3.5</td>
<td>8</td>
</tr>
<tr>
<td>4-5</td>
<td>4.5</td>
<td>11</td>
</tr>
<tr>
<td>5-6</td>
<td>5.5</td>
<td>13</td>
</tr>
<tr>
<td>6-7</td>
<td>6.5</td>
<td>12</td>
</tr>
<tr>
<td>7-8</td>
<td>7.5</td>
<td>8</td>
</tr>
<tr>
<td>8-9</td>
<td>8.5</td>
<td>9</td>
</tr>
<tr>
<td>9-10</td>
<td>9.5</td>
<td>5</td>
</tr>
</tbody>
</table>

The average life of insects can be calculated by the formula as

$$\bar{x} = \frac{3 \times 0.5 + 9 \times 1.5 + \ldots + 5 \times 9.5}{3 + 9 + \ldots + 5}$$

$$= \frac{431}{80}$$

$$= 5.39 \text{ days}$$

**Weighted mean:**

In case, $k$-variate values $X_1, X_2, \ldots, X_k$ have known weights $\omega_1, \omega_2, \ldots, \omega_k$ respectively, then the weighted mean is

$$\mu = \frac{\omega_1 X_1 + \omega_2 X_2 + \ldots + \omega_k X_k}{\omega_1 + \omega_2 + \ldots + \omega_k}$$

$$= \frac{1}{\omega} \sum_i \omega_i X_i$$
Where $\omega = \sum \omega_i$, where $i = 1, 2, \ldots k$

Weighted mean is commonly used in the construction of index numbers.

*Note:* If the sample values $x_1, x_2, \ldots, x_n$ are given, in the formulae of mean given above, capital $X$ will be changed in small $x$ and $N$ to $n$. The sample mean will be denoted by $x$ e.g. the formula $\bar{x}$ will be changed to

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$= \frac{1}{n} \sum_{i=1}^{n} x_i$$

$i = 1, 2, \ldots, n$

Similarly, all other formulae will be changed to sample values. Since most of the studies are based on samples in practice, we use formulae for sample values.

**Merits and demerits (A.M.)**

1. Algebraic sum of deviations of the given values for their arithmetic mean is always zero i.e. $\sum (X_i - \bar{X}) = 0$.

2. The sum of squares of the deviations of the given values from their A.M. is minimum, i.e. $\sum (X_i - \bar{X})^2$ is minimum.

3. An average possesses all the characteristics of a central value given earlier except No. 2, which is greatly affected by the extreme values.

4. In case of grouped data if any class interval is open, arithmetic mean cannot be calculated, e.g. the classes are less than five in the beginning or more than 70 at the end of the distribution or both.

**Example:** Daily cash earnings (in Rs.) of 15 laborers engaged in different field works are as follows:

11.63, 8.22, 12.56, 12.14, 29.23, 18.23, 11.49, 11.30,
17.00, 9.16, 8.64, 27.56, 8.23, 19.77, 12.81,

The average daily earning of a worker can be calculated by the formula as

$$A = \frac{1}{15} (11.63 + 8.22 + \ldots + 12.81)$$

$$= \frac{217.97}{15}$$

$$= 14.53$$

The average daily earning of a worker is Rs.14.53
Example: The distribution of the size of the holding of cultivated land, in an area was as follows

<table>
<thead>
<tr>
<th>Size of holding (Ha)</th>
<th>Mid points (y)</th>
<th>No. of holdings (f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-2</td>
<td>1</td>
<td>48</td>
</tr>
<tr>
<td>2-4</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>4-6</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>6-8</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>8-10</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>10-20</td>
<td>15</td>
<td>9</td>
</tr>
<tr>
<td>20-40</td>
<td>30</td>
<td>2</td>
</tr>
<tr>
<td>40-60</td>
<td>50</td>
<td>1</td>
</tr>
</tbody>
</table>

The average size of holding in the area can be calculated with the help of the formula. Mid points of the class intervals are shown in the middle column along with the data. Hence:

$$A.M. = \frac{1 \times 48 + 3 \times 19 + \ldots + 50 \times 1}{48+19+\ldots+1}$$

$$= \frac{597}{114}$$

$$= 5.237$$

The average size of holding is 5.237 hectares.

Median

Median of a set of values is the value which is the middle most value when they are arranged in the ascending order of magnitude. Median is denoted by ‘M’. In case of discrete series without or with frequency, it is given by:

$$M = \left( \frac{n+1}{2} \right)^{th}$$

To solve the problems on median:
1. Arrange the data in ascending order or descending order.
2. Make class-interval as exclusive type.

**Example:** Find the median value of the following set of values
45, 32, 31, 46, 40, 28, 27, 37, 36, 41, 47, 50.

Arranging in ascending order we get:
27, 28, 31, 32, 36, 37, 40, 41, 45, 46, 47, 50

We have, \( n = 12 \)

The median for the given set of values is 38.5.

**Mode**

Mode is the value which has the highest frequency and is denoted by \( Z \).

**Example:** The following data relate to size of shoes. Find the mode.
6, 7, 6, 8, 9, 9, 10, 8, 7, 7, 9, 10, 9, 9, 9, 8, 8, 11

**Solution:** Arrange the data in ascending order, data obtained is shown in the following Table.

**Table:** Frequency table for data in solved problem.

<table>
<thead>
<tr>
<th>Size</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
</tr>
</tbody>
</table>

Therefore, mode value is 9, which is corresponding to the highest frequency 7.
Measures of Dispersion

Median is a positional average and has nothing to do with the variability of the observations in the series, Mode is the largest occurring value independent of other values of the set. This leads us to conclude that a measure of central tendency alone is not enough to have a clear idea about the data unless all observations are almost the same. Moreover, two or more sets may have the same mean and/or median but they may be quite different.

A measure of dispersion or variability observed the scatteredness amongst observations in the set. Commonly used measures of dispersion are:

1. Range
2. Interquartile range and Quartile deviation
3. Mean deviation
4. Variance
5. Standard deviation
6. Coefficient of variation

Range

Definition: It is the difference between the largest and smallest observation in a set.

If we denote the largest observation by \( L \) and the smallest observation by \( S \), the formula is,

\[
\text{Range } R = L - S
\]

A relative measure known as coefficient of range is given as,

\[
\text{Coefficient of range } = \frac{L - S}{L + S}
\]

Lesser the range or coefficient of range, better the result.

Properties

1. It is the simplest measure and can easily be understood.
2. Besides the above merit, it hardly satisfies any property of a good measure of dispersion e.g. it is based on two extreme values only, ignoring the others. It is not liable to further algebraic treatment.
3. Example: The plant height (cm) of eighteen plants of a crop is as given below:

\[
\begin{align*}
77, & \quad 76, & \quad 83, & \quad 68, & \quad 57, & \quad 107, & \quad 80, & \quad 75, & \quad 95 \\
100, & \quad 113, & \quad 119, & \quad 121, & \quad 121, & \quad 83, & \quad 87, & \quad 46, & \quad 74
\end{align*}
\]
We can find the range by the formula given above

\[ L = 121 \text{ and } S = 46 \]

The range,

\[ R = 121 - 46 = 75 \]

Some people also write the range as 46-121.

Also,

\[ \text{Coefficient of range} = \frac{121 - 46}{121 + 46} = \frac{75}{167} = 0.449 \]

**Interquartile range (I.R.)**

*Definition*: The difference between the third quartile and first quartile is called interquartile range. Symbolically

\[ \text{I.R.} = Q_3 - Q_1 \]

**Quartile deviation (Q.D.)**

This is half of the interquartile range i.e.

\[ \text{Q.D.} = \frac{Q_3 - Q_1}{2} \]

Also the coefficient of quartile deviation is given by the formula

\[ \text{Coeff. of Deviation} = \frac{Q_3 - Q_1}{Q_3 - Q_1} \]

Coefficient of quartile deviation is an absolute quantity (unit less) and is useful to compare the variability among the middle 50% observations.

**Properties**

(1) It is better measure of dispersion than range in the sense that it involves 50% of the mid values of a series of data rather than only two extreme values of a series.

(2) Since it excludes the lowest and highest 25% values, it is not affected by the extreme values.

(3) It can be calculated for the grouped data with open end intervals.

(4) It is not capable of further algebraic treatment.

(5) It is susceptible to sampling fluctuations.

Example: The value of Q1, Q2 and Q3 are \( Q_1 = 174.90; Q_2 = 190.23; Q_3 = 203.83 \).

Interquartile range, \( \text{I.R.} = 203.83 - 174.90 = 28.93 \).

Quartile deviation, \( \text{Q.D.} = \frac{203.83 - 174.90}{2} = \frac{28.93}{2} = 14.465 \)

Coeff. of \( \text{Q.D.} = \frac{203.83 - 174.90}{203.83 + 174.90} = \frac{28.93}{378.73} = 0.076 \)
Mean deviation (M.D.)

The measures of dispersion discussed so far are not satisfactory in the sense that they lack most of the requirements of a good measure. Mean deviation is a better measure than range and Q.D.

*Definition:* It is the average of the absolute deviations taken from a central value, generally the mean or median.

Considering a set of \( N \) observations \( X_1, X_2, \ldots, X_N \). Then the mean deviation

\[
M.D. = \frac{1}{N} \sum_{i=1}^{N} |X_i - A|
\]

*Properties:*

1. Mean deviations remove one main objection of the earlier measures, that it involves each value of the set.
2. It is not affected much by extreme values.
3. It has no relationship with any of the other measures of dispersion.
4. Its main drawback is that algebraic negative signs of the deviations are ignored which is mathematically unsound.
5. Mean deviation is minimum when the deviations are taken from median.

*Example:* The total cereal production of all crops in India from 1991 to 1998 is given below:

Production (million tonnes)

111.5, 111.2, 102.3, 112.4, 108.8, 125.3, 116.5, 132.7

Mean deviation taking deviation from the mean and also from the median has been calculated.

(i) Mean = \( \frac{1}{8} (111.5 + 111.2 + \ldots + 132.7) = \frac{920.7}{8} = 115.09 \)

Mean deviation taking deviations from the mean using the formula is

\[
M.D. = \frac{1}{8} \left( |111.5 - 115.09| + |111.2 - 115.09| + \ldots + |132.7 - 115.09| \right)
\]

\[
= \frac{1}{8} (3.59 + 3.89 + 12.79 + 2.69 + 6.29 + 10.21 + 1.41 + 17.61)
\]

\[
= \frac{58.48}{8} = 7.31 \text{ million tonnes.}
\]
(ii) Calculate the mean deviation taking the deviations from the median. For finding out the median, arrange the data in ascending order:

102.3, 108.8, 111.2, 111.4, 112.4, 116.5, 125.3, 132.7

Median = \frac{115.5 + 112.4}{2} = \frac{229.9}{2} = 111.95

Mean deviation taking deviations from the median by the formula

\text{M.D.} = \frac{1}{8} \left( |102.3 - 111.95| + |108.8 - 111.95| + \ldots + |132.7 - 111.95| \right)

\text{M.D.} = \frac{1}{8} (9.65 + 3.15 + 0.75 + 0.45 + 0.45 + 4.55 + 13.35 + 20.75)

\text{M.D.} = \frac{53.1}{8} = 6.64 \text{ million tonnes.}

\textbf{Example:} Distribution of tuber weight (in g) of range of 15-39 is displayed here.

<table>
<thead>
<tr>
<th>Tuber weight (in g)</th>
<th>15-19</th>
<th>19-23</th>
<th>23-27</th>
<th>27-31</th>
<th>31-35</th>
<th>35-39</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of tubers</td>
<td>8</td>
<td>59</td>
<td>47</td>
<td>23</td>
<td>6</td>
<td>4</td>
</tr>
</tbody>
</table>

Mean deviation taking deviations from the mean has been calculated. The calculations are shown in the Table given below:

| Class intervals | Midpoints (X) | Frequency (f) | fx     | \(|X - \overline{X}|\) \(f \left|X - \overline{X}\right|\) |
|-----------------|--------------|-------------|-------|----------------------|
| 15-19           | 17           | 8           | 136   | 7.24                 | 57.92   |
| 19-23           | 21           | 59          | 1239  | 3.24                 | 191.16  |
| 23-27           | 25           | 47          | 1175  | 0.76                 | 35.72   |
| 27-31           | 29           | 23          | 667   | 4.76                 | 109.48  |
| 31-35           | 37           | 6           | 198   | 8.76                 | 52.56   |
| 35-39           | 33           | 4           | 148   | 12.76                | 51.04   |
| Total           | 147          | 3563        |       |                       | 407.88  |

\[ \text{where, } X = \frac{3563}{147} = 24.24 \]

Mean deviation about mean by the formula is
M.D. $= \frac{497.88}{147} = 3.39 \text{ g}$

Variance:

The main objection of mean deviation is that the negative signs are ignored, is removed by taking the square of the deviations from the mean.

**Definition:** The variance is the average of the squares of the deviations taken from mean.

Let $X_1, X_2, \ldots, X_N$ be the measurements on $N$ population units, the population variance.

$$\sigma^2 = \frac{1}{N} \sum_{i} (X_i - \bar{X})^2$$

for $i = 1, 2, 3, \ldots, N$ and $\bar{X}$ is the population mean.

The sample variance of the set $x_1, x_2, \ldots x_n$ of $n$ observations is given by the formula.

$$s^2 = \frac{1}{n-1} \sum_{i} (x_i - \bar{x})^2$$

$$= \frac{1}{n-1} \left( \sum_{i} x_i^2 - \left( \sum_{i} x_i \right)^2 / n \right)$$

for $i = 1, 2, 3, \ldots, n$, where $\bar{x} = \frac{1}{n} \sum_{i} X_i$.

If the observation $x_i$ occurs $f_i$ times for $i = 1, 2, \ldots, k$, then the sample variance

$$s^2 = \frac{1}{n-1} \sum_{i} f_i (x_i - \bar{x})^2$$

$$= \frac{1}{n-1} \left( \sum_{i} f_i x_i^2 - \left( \sum_{i} f_i x_i \right)^2 / n \right)$$

Where, $n = \sum f_i$.

**Example:** The following nine measurements are the heights in inch in a sample of nine plants.

| Height (X): 69, 66, 67, 69, 64, 63, 65, 68, 72 |

The sample variance of height of soldiers can be computed by formula

$$\sum_{i=1}^{9} X_i = 69 + 66 + \ldots + 72 = 603$$
Standard Deviation and Root Mean Square Deviation

Standard deviation is the positive square root of the arithmetic mean of the square of the deviations of the given values from their arithmetic mean. It is denoted by Greek letter $\sigma$ (sigma). Mathematically the standard deviation is expressed as

$$
\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}
$$

(for ungrouped data)

$$
\sigma = \sqrt{\frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^2}{\sum_{i=1}^{n} f_i}}
$$

(for frequency distribution)

Where, $X_i$ is the variable; $\bar{X}$ is the arithmetic mean; $n$ or $\sum_{i=1}^{n} f_i$ is the total number of observations.

The step of squaring the deviation $(X_i - \bar{X})^2$ overcomes the drawback of ignoring the sign in mean deviation.

The square of standard deviation ($\sigma^2$) is called the variance.

In certain respects variance is even more important than the standard deviation. Moreover, of all the measures, standard deviation is affected least by fluctuations of sampling.

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

**Example**  TSS levels of 10 varieties of a fruit crop are as under:

240, 260, 290, 245, 288, 272, 263, 277, 251

Calculate standard deviation with the help of assumed mean.

**Solution** Calculation of standard deviation by the assumed mean method

<table>
<thead>
<tr>
<th>X</th>
<th>d=(x – 264)</th>
<th>d²</th>
</tr>
</thead>
<tbody>
<tr>
<td>240</td>
<td>-24</td>
<td>576</td>
</tr>
<tr>
<td>260</td>
<td>-4</td>
<td>16</td>
</tr>
<tr>
<td>290</td>
<td>+26</td>
<td>676</td>
</tr>
<tr>
<td>245</td>
<td>-19</td>
<td>361</td>
</tr>
<tr>
<td>255</td>
<td>-9</td>
<td>81</td>
</tr>
<tr>
<td>288</td>
<td>+24</td>
<td>576</td>
</tr>
<tr>
<td>272</td>
<td>+8</td>
<td>64</td>
</tr>
<tr>
<td>263</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>277</td>
<td>+13</td>
<td>169</td>
</tr>
<tr>
<td>251</td>
<td>-13</td>
<td>169</td>
</tr>
</tbody>
</table>

\[
\sum X = 2641 \quad \sum d = +1 \quad \sum d^2 = 2689
\]

\[
\sigma = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2}
\]

\[
\sum d^2 = 2689, \quad \sum d = +1 \quad N=10
\]

\[
\sigma = \sqrt{\frac{2689}{10} - \left(\frac{1}{10}\right)^2} = \sqrt{268.9 - 0.01} = 16.398
\]

**Properties of standard deviation**

1. Standard deviation is independent of change of origin.

   Shortcut method of computing standard deviation proves this property. There we are changing the origin by shifting the origin of the variable by the point \( A \).

2. Standard deviation is not independent of change of scale.

   Step deviation method of computing standard deviation proves this property. There we are changing the scale first by dividing the variable by a common factor \( i \) and then to get the original standard deviation it is again multiplied by that common factor \( (i) \).
3. When two distributions are combined into a single distribution it is possible to
calculate the standard deviation of the combined series.

If \( \bar{x}_1 \) and \( \bar{x}_2 \) are the means of two series of size \( n_1 \) and \( n_2 \) with variance \( \sigma_1^2 \) and \( \sigma_2^2 \), respectively. Then the formula of variance of the series formed by adding the two
given series is given by

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

where \( d_1^2 = (\bar{x}_1 - \bar{x})^2 \), \( d_2^2 = (\bar{x}_2 - \bar{x})^2 \) and \( \bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} \)

**Coefficient of Dispersion**

Whenever we want to compare the variability of the two series which differ
widely an their averages or which are measured in different units, we do not merely
calculate the measures of dispersion but we calculate the coefficients of dispersion Which
are pure numbers independent of the units of measurement. The coefficients of dispersion
(C.D.) based on different measures of dispersion are as follows:

1. C.D. based upon range \( = \frac{A - B}{A + B} \)

   where A and B are the greatest and the smallest items in the series.

2. Based upon quartile deviation \( \frac{(Q_3 - Q_1)/2}{(Q_3 - Q_1)/2} = \frac{(Q_3 - Q_1)}{(Q_3 - Q_1)} \)

3. based upon mean deviation \( = \frac{\text{mean deviation}}{\text{average from which it is calculated}} \)

4. Based upon standard deviation \( C.D. = \frac{S.D.}{Mean} = \frac{\sigma}{\bar{x}} \)

**Coefficient of Variation:** 100 times the coefficient of dispersion based upon standard
deviation is called coefficient of variation (C.V.), i.e.,

\[
\text{C.V.} = 100 \times \frac{\sigma}{\bar{x}}
\]

According to Professor Karl Pearson who suggested this measure, C.V. is the percentage
variation in the mean, standard deviation being considered as the total variation in the
mean.

For comparing the variability of two series, we calculate the coefficient of variations for
each series. the series having greater CV. is said to be more variable than the other and
the series having lesser C.V. is said to be more consistent (or homogeneous) than the other.

**Example**

An analysis of monthly wages paid to the laborers of two commercial farms A and B belonging to the same crop gives the following results.

<table>
<thead>
<tr>
<th></th>
<th>Farm A</th>
<th>Farm B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of laborers</td>
<td>500</td>
<td>600</td>
</tr>
<tr>
<td>Average daily wage</td>
<td>Rs. 78600</td>
<td>Rs. 17500</td>
</tr>
<tr>
<td>Variance of distribution of wages</td>
<td>81</td>
<td>100</td>
</tr>
</tbody>
</table>

(i) Which farm A or B, has a larger wage bill?

(ii) In which farm, A or B, is there greater variability in individual wages

(iii) Calculate (a) the average daily range, and (b) the variance of the distribution of wages of all the laborers in the farm A and B taken together.

**Solution**

(i)

Farm A:

No. of laborers, (say), \( n_1 = 500 \), Average daily wages, (say), \( \bar{x}_1 = Rs. 186 \)

Average daily wage = \( \frac{\text{Total wages paid}}{\text{No. of workers}} \)

Hence total wages paid to the workers = \( n_1 \bar{x}_1 = 500 \times 186 = Rs. 93,000 \)

Farm B:

No. of laborers, (say), \( n_2 = 600 \); Average daily wages, (say), \( \bar{x}_2 = Rs. 175 \)

Total daily wages paid to the workers = \( n_2 \bar{x}_2 = 600 \times 175 = Rs. 1,05,000 \)

Thus we see that the farm B has larger wage bill.

(ii)

Variance of distribution of wages in farm A, (say), \( \sigma_1^2 = 81 \)

Variance of distribution of wages in farm B, (say), \( \sigma_2^2 = 100 \)

C.V. of distribution of wages for farm A = \( 100 \times \frac{\sigma_1}{\bar{x}_1} = \frac{100 \times 9}{186} = 4.84 \)

CV. of distribution of wages for farm B = \( 100 \times \frac{\sigma_2}{\bar{x}_2} = \frac{100 \times 10}{175} = 5.71 \)

Since CV. for farm B is greater than CV. for farm A, farm B has greater variability in individual wages.

(i)

(a) The average daily wages (say) \( \bar{x} \), of all the laborers in the two farms A and B taken together is given by:
\[ \bar{x} = \frac{x_1 x_2 + x_2 x_2}{n_1 + n_2} = \frac{500x186 + 600x175}{500 + 600} = \frac{1,98,000}{1,100} = \text{Rs. 180} \]

(a) The combined variance \( \sigma^2 \) is given by the formula:

\[ \sigma^2 = \frac{1}{n_1 + n_2} \left[ n_1 (\sigma_1^2 + d_1^2) + n_2 (\sigma_2^2 + d_2^2) \right], \text{ where } d_1 = x_1 - \bar{x} \text{ and } d_2 = x_2 - \bar{x}. \]

Here \( d_1 = 186 - 180 = 6 \) and \( d_2 = 175 - 180 = 5 \)

Hence \( \sigma^2 = \frac{500(81 + 36) + 600(100 + 25)}{500 + 600} = \frac{1,33,500}{1,100} = \text{Rs. 121.36} \)

**SKEWNESS AND KURTOSIS**

There are two other comparable characteristics called skewness and kurtosis that help us to understand a distribution.

Definition

“Skewness refers to the asymmetry or lack of symmetry in the shape of frequency distribution.”

**Difference between Dispersion and Skewness**

Dispersion is concerned with the amount of variation rather than with its direction. Skewness tells us about the direction of the variation or the departure from symmetry. In fact, measures of skewness are dependent upon the amount of dispersion.

It may be noted that although skewness is an important characteristic for defining the precise pattern of a distribution, it is rarely calculated in business and economic series. Variation is by far the most important characteristic of a distribution.

**Karl Pearson’s Coefficient of Skewness**

This method of measuring skewness, also known as Pearsonian Coefficient of Skewness, was suggested by Karl Pearson\(^*\), a great British Biometrician and Statistician. It is based upon the difference between mean and mode. This difference is divided by standard deviation to give a relative measure. The formula thus becomes

\[ S_{kp} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}} \]

\( S_{kp} = \text{Karl Pearson’s coefficient of skewness.} \)
There is no limit to this measure in theory and this is a slight drawback. But in practice the value given by this formula is rarely very high and usually lies between ± 1.

When a distribution is symmetrical, the values of mean, median and mode coincide and, therefore, the coefficient of skewness will be zero. When a distribution is positively skewed, the coefficient of skewness shall have plus sign and when it is negatively skewed, the coefficient of skewness shall have minus sign. The degree of skewness shall be obtained by the numeral value. say, 0.8 or 0.2 etc. Thus, this formula gives both the direction as well as the extent of skewness.

The above method of measuring skewness cannot be used where mode is ill defined. However, in moderately skewed distribution the averages have the following relationship

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

and therefore, if this value of mode is substituted in the above formula we arrive at another formula for finding out skewness,

\[ S_{kp} = \frac{3(\text{Mean} - \text{Median})}{\sigma} \]

Theoretically, the value of this coefficient varies between ± 3

**Example** Calculate Karl Pearson’s coefficient of skewness:

<table>
<thead>
<tr>
<th>X</th>
<th>F</th>
<th>(d=(x-27.5)/5)</th>
<th>(fd)</th>
<th>(fd^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.5</td>
<td>28</td>
<td>-3</td>
<td>-84</td>
<td>252</td>
</tr>
<tr>
<td>17.5</td>
<td>42</td>
<td>-2</td>
<td>-84</td>
<td>168</td>
</tr>
<tr>
<td>22.5</td>
<td>54</td>
<td>-1</td>
<td>-54</td>
<td>54</td>
</tr>
<tr>
<td>27.5</td>
<td>108</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>32.5</td>
<td>129</td>
<td>+1</td>
<td>+129</td>
<td>129</td>
</tr>
<tr>
<td>37.5</td>
<td>61</td>
<td>+2</td>
<td>+122</td>
<td>244</td>
</tr>
<tr>
<td>42.5</td>
<td>45</td>
<td>+3</td>
<td>+135</td>
<td>405</td>
</tr>
<tr>
<td>47.5</td>
<td>33</td>
<td>+4</td>
<td>+132</td>
<td>528</td>
</tr>
</tbody>
</table>

\[ N = 500 \quad \Sigma fd = 296 \quad \Sigma fd^2 = 1780 \]

Coeff. of Sk. = \( \frac{\text{Mean} - \text{Mode}}{\sigma} \)
\[
\bar{x} = A + \frac{\sum fd}{d} \times i
\]

A = 27.5, \(\Sigma fd = 296\), \(N = 500\), \(I = 5\).

\[
\bar{x} = 27.5 + \frac{296}{500} \times 5 = 30.46
\]

**Mode** : since the maximum frequency is 129, the corresponding value of \(X\), i.e., 32.5 is the modal value.

S.D. \(\sigma = i \times \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2}
\]

\(\Sigma fd^2 = 1780\), \(N = 500\), \(\Sigma fd = 296\), \(i = 5\)

\[
\sigma = 5 \times \sqrt{\frac{1780}{500} - \left(\frac{296}{500}\right)^2} = 5 \times \sqrt{3.56 - 0.35} = 8.96
\]

Coeff. of Sk. \(\frac{30.46 - 32.5}{8.96} = -0.228\)

**Example**  Calculate Karl Pearson’s coefficient of Skewness from the following data:

<table>
<thead>
<tr>
<th>Profits (Rs. Lakhs)</th>
<th>No. of cos.</th>
<th>Profits (Rs. Lakhs)</th>
<th>No. of cos.</th>
</tr>
</thead>
<tbody>
<tr>
<td>70-80</td>
<td>12</td>
<td>110-120</td>
<td>50</td>
</tr>
<tr>
<td>80-90</td>
<td>18</td>
<td>120-130</td>
<td>45</td>
</tr>
<tr>
<td>90-100</td>
<td>35</td>
<td>130-140</td>
<td>30</td>
</tr>
<tr>
<td>100-110</td>
<td>42</td>
<td>140-150</td>
<td>8</td>
</tr>
</tbody>
</table>

**Solution** Calculation of coeff. of skewness by Karl Pearson’s method

<table>
<thead>
<tr>
<th>Profits. (Rs. Lakhs)</th>
<th>(m - 115)/10</th>
<th>fd</th>
<th>Fd^2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(m)</td>
<td>f</td>
<td>d</td>
</tr>
<tr>
<td>70-80</td>
<td>75</td>
<td>12</td>
<td>-4</td>
</tr>
<tr>
<td>80-90</td>
<td>85</td>
<td>18</td>
<td>-</td>
</tr>
<tr>
<td>90-100</td>
<td>95</td>
<td>35</td>
<td>-2</td>
</tr>
<tr>
<td>100-110</td>
<td>105</td>
<td>42</td>
<td>-1</td>
</tr>
<tr>
<td>110-120</td>
<td>115</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>120-130</td>
<td>125</td>
<td>45</td>
<td>+1</td>
</tr>
<tr>
<td>130-140</td>
<td>135</td>
<td>30</td>
<td>+2</td>
</tr>
<tr>
<td>140-150</td>
<td>145</td>
<td>8</td>
<td>+3</td>
</tr>
</tbody>
</table>

\(N = 240\) \(\Sigma fd = 85\) \(\Sigma fd^2 = 773\)
Coeff. of Sk. = \( \frac{\text{Mean} - \text{Mode}}{\sigma} \)

Mean : \( \bar{x} = A + \frac{\sum fd}{d} \times i = 115 - \frac{85}{240} \times 10 = 115 - 111.46 \)

Mode : by inspection mode lies in the class 110 -120.

Mode = \( \angle + \frac{\Delta l}{\Delta 1 + \Delta 2} \times i \)

\( \angle = 110, \Delta l = 150-421 =8, \Delta 2150-451=5, i =10 \)

Mode = 110 + \( \frac{8}{8+5} \times 10 = 110 + 6.15 = 116.15 \)

S.D. \( \sigma = i \times \sqrt{\frac{\sum d^2}{N} - \left( \frac{\sum d}{N} \right)^2} = 10 \times \sqrt{\frac{773}{240} - \left( \frac{85}{240} \right)^2} \)

= \( \sqrt{3.221 - 0.125} \times 10 = 1.7595 \times 10 = 17.595 \)

Coeff. of sk. \( \frac{111.46 - 116.15}{17.595} = -\frac{4.69}{17.595} = -0.266 \)

**Measures of Kurtosis**

The most important measure of kurtosis is the value of the coefficient \( \beta_2 \) It is defined as:

\( \beta_2 = \frac{\mu_4}{\mu_2^2} \) where \( \mu_4 = 4\text{th moment} \) and \( \mu_2 = 2\text{nd moment} \).

The greater the value of \( \beta_2 \) the more peaked is the distribution.

For a normal curve, the value of \( \beta_2 = 3 \). When the value of \( \beta_2 \) is greater than 3, the curve is more peaked than the normal curve, i.e., leptokurtic. When the value of \( \beta_2 \) less than 3 then the curve is less peaked than the normal curve, i.e., platykurtic. The normal curve and other curves with \( \beta_2 = 3 \) are called mesokurtic.

Sometimes \( Y_2 \) the derivative of \( \beta_2 \) is used as a measure of kurtosis. \( Y_2 \) is defined as \( Y_2 = \beta_2 - 3 \)

For a normal distribution, \( Y_2 = 0 \). If \( Y_2 \) is positive, the curve is leptokurtic and if \( Y_2 \) is negative, the curve is platykurtic.
Example Compute the coefficient of skewness and kurtosis based on the following data:

<table>
<thead>
<tr>
<th>X</th>
<th>f</th>
<th>d = (x – 44.5) /10</th>
<th>fd</th>
<th>fd^2</th>
<th>fd^3</th>
<th>fd^4</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.5</td>
<td>1</td>
<td>-4</td>
<td>-4</td>
<td>16</td>
<td>-64</td>
<td>256</td>
</tr>
<tr>
<td>14.5</td>
<td>5</td>
<td>-3</td>
<td>-15</td>
<td>45</td>
<td>-135</td>
<td>405</td>
</tr>
<tr>
<td>24.5</td>
<td>12</td>
<td>-2</td>
<td>-24</td>
<td>48</td>
<td>-96</td>
<td>192</td>
</tr>
<tr>
<td>34.5</td>
<td>22</td>
<td>-1</td>
<td>-22</td>
<td>22</td>
<td>-22</td>
<td>22</td>
</tr>
<tr>
<td>44.5</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>54.5</td>
<td>9</td>
<td>+1</td>
<td>+9</td>
<td>9</td>
<td>+9</td>
<td>9</td>
</tr>
<tr>
<td>64.5</td>
<td>4</td>
<td>+2</td>
<td>+8</td>
<td>16</td>
<td>+32</td>
<td>64</td>
</tr>
<tr>
<td>74.5</td>
<td>3</td>
<td>+3</td>
<td>+9</td>
<td>27</td>
<td>+81</td>
<td>243</td>
</tr>
<tr>
<td>84.5</td>
<td>1</td>
<td>+4</td>
<td>+4</td>
<td>16</td>
<td>+64</td>
<td>256</td>
</tr>
<tr>
<td>94.5</td>
<td>1</td>
<td>+5</td>
<td>+5</td>
<td>25</td>
<td>+125</td>
<td>625</td>
</tr>
</tbody>
</table>

N = 75 \quad \Sigma fd = 30 \quad \Sigma fd^2 = 224 \quad \Sigma fd^3 = -6 \quad \Sigma fd^4 = 2,072

\[ \mu_1 = \frac{\Sigma fd}{N} = \frac{-30}{75} = -0.4; \quad \mu_2 \ast = \frac{\Sigma fd^2}{N} = \frac{224}{75} = -2.99; \]

\[ \mu_3 \ast = \frac{\Sigma fd^3}{N} = \frac{-6}{75} = -0.8; \quad \mu_4 = \frac{\Sigma fd^4}{N} = \frac{2,072}{75} = -27.63; \]

\[ \mu_2 \ast = \mu_2 - (\mu_2)^2 = 2.99 - (-0.4)^2 = 2.99 - 0.16 = 2.83 \]

\[ \mu_3 \ast = \mu_3 - 3\mu_1 \mu_2 + 2(\mu_1)^3 = -0.08 - 3(-0.4)(2.99) + 2(-0.4)^3 \]

\[ = -0.08 + 3.588 - 0.128 = 3.38 \]

\[ \mu_4 \ast = \mu_4 - 4\mu_1 \mu_3 + 6\mu_1^2 \mu_2 - 3(\mu_1)^4 \]

\[ = 27.63 - 4(-0.4)(-0.08) + 6(-0.4)^2(2.99) + 3(-0.4)^4 \]

\[ = 27.63 + 0.128 + 2.87 - 0.077 = 30.295 \]

\[ \beta_1 \ast = \frac{\mu_3^2}{\mu_2} = \frac{(3.38)^2}{(2.83)^3} = \frac{11.424}{22.665} = 0.504. \]

For kurtosis we have to compute the value of \( \beta_2 \)

\[ \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{30.295}{(2.83)^2} = \frac{30.295}{8.01} = 3.782 \]
**Example**  The first four central moments of distribution are 0, 2.5, 0.7 and 18.75 comment on the skewness and kurtosis of the distribution.

**Solution**  Testing skewness

We are given \( \mu_2 = 2.5, \mu_3 = 0.7 \) and \( \mu_4 = 18.75 \)

Skewness is measured by the coefficient \( \beta_1 \)

\[
\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{0.7}{(2.5)^3} = 0.031
\]

Since \( \beta_1 = 0.031 \), the distribution is slightly skewed, i.e., it is not perfectly symmetrical.

For testing kurtosis we compute the value of \( \beta_2 \) when a distribution is normal or symmetrical, \( \beta_2 = 3 \). When distribution is more peaked than the normal, \( \beta_2 \) is more than 3 and when it is less peaked than the normal \( \beta_2 \) is less than 3.

\[
\beta_2 = \frac{\mu_4^2}{\mu_2^4}; \text{ where } \mu_4 = 18.75, \mu_2 = 2.5
\]

\[
\beta_2 = \frac{18.75^2}{(2.5)^4} = \frac{18.75}{6.25} = 3
\]

Since \( \beta_2 \) is exactly three, the distribution is mesokurtic.

**Self Assessment Questions**

**State whether the following questions are ‘True’ or ‘False’**.

**SET-1**

i. For a given set of values if we add a constant 5 to every value, then the arithmetic mean is affected.

ii. Arithmethic mean can be calculated for distribution with open-end classes.

iii. Arithmetic mean is affected by extreme values.

iv. Arithmetic mean of 12, 16, 23, 25, 28, 32 is 22.

**SET-2**

i. Mode is based on all values

ii. Mode = 3 Median – Mean

iii. Geometric mean is used when we are interested in rate of growth of any phenomenon.

iv. Harmonic mean exists if one of the values is zero

v. A.M. < G.M. < H.M. for any two value ‘a’ and ‘b’.
vi. Arithmetic mean can be calculated accurately even when the distribution has
open end class.

vii. Mode can be located graphically.

viii. Mode is used when data is on interval scale.

ix. Standard deviation is based on all values

x. Standard deviation of a set of values is increased if every value of the set is
increased by a constant.

xi. Standard deviation can be calculated for distributions with open-end classes.

Unsolved questions

1. In an office there are 84 employees. Their salaries in Indian rupees are as given in Table below. Find the mean salary per day

<table>
<thead>
<tr>
<th>Salary/day</th>
<th>Employees</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>3</td>
</tr>
<tr>
<td>70</td>
<td>5</td>
</tr>
<tr>
<td>80</td>
<td>8</td>
</tr>
<tr>
<td>90</td>
<td>10</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
</tr>
<tr>
<td>120</td>
<td>2</td>
</tr>
</tbody>
</table>

2. A survey of 128 smokers gave the results represented in Table below, which are
frequency distribution of smokers’ daily expenses on smoking. Find the mean expenses
and standard deviation. Determine coefficient of variation.

<table>
<thead>
<tr>
<th>Expenditure (Rs.)</th>
<th>No. of Smokers</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-20</td>
<td>23</td>
</tr>
<tr>
<td>20-30</td>
<td>44</td>
</tr>
<tr>
<td>30-40</td>
<td>35</td>
</tr>
<tr>
<td>40-50</td>
<td>12</td>
</tr>
<tr>
<td>50-60</td>
<td>9</td>
</tr>
<tr>
<td>60-70</td>
<td>3</td>
</tr>
<tr>
<td>70-80</td>
<td>2</td>
</tr>
</tbody>
</table>

3. The average price /kg of Grade ‘A’ tea is Rs.120 and that of grade ‘B’ tea is Rs.100. A
trader mixes then and sell the mixture for Rs.115. Find proportion of Grave A and grade
B in the mixture.

4. For the distribution shown in Table, find the median and mode.

<table>
<thead>
<tr>
<th>% marks</th>
<th>No. of Smokers</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-10</td>
<td>4</td>
</tr>
<tr>
<td>10-20</td>
<td>9</td>
</tr>
<tr>
<td>20-30</td>
<td>19</td>
</tr>
<tr>
<td>30-40</td>
<td>20</td>
</tr>
<tr>
<td>40-50</td>
<td>18</td>
</tr>
<tr>
<td>50-60</td>
<td>7</td>
</tr>
<tr>
<td>60-70</td>
<td>80</td>
</tr>
</tbody>
</table>
5. Find the geometric mean of the following distribution.

<table>
<thead>
<tr>
<th>X</th>
<th>110</th>
<th>115</th>
<th>118</th>
<th>119</th>
<th>120</th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>4</td>
<td>11</td>
<td>21</td>
<td>6</td>
<td>2</td>
</tr>
</tbody>
</table>

6. Find the harmonic mean of the following distribution.

<table>
<thead>
<tr>
<th>X</th>
<th>121</th>
<th>122</th>
<th>123</th>
<th>124</th>
<th>125</th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>5</td>
<td>25</td>
<td>36</td>
<td>37</td>
<td>20</td>
</tr>
</tbody>
</table>

7. Find the quartile deviation and the coefficient of quartile deviation for the data shown in table.

<table>
<thead>
<tr>
<th>Age group</th>
<th>15-20</th>
<th>20-25</th>
<th>25-30</th>
<th>30-35</th>
<th>35-40</th>
<th>40-45</th>
<th>Above 45</th>
</tr>
</thead>
<tbody>
<tr>
<td>% of people who exercise regularly</td>
<td>15</td>
<td>31</td>
<td>19</td>
<td>15</td>
<td>8</td>
<td>7</td>
<td>7</td>
</tr>
</tbody>
</table>

8. Given sum of upper and lower quartiles as 122 and their difference as 23; find the quartile deviation of the series.

9. If C.V. % = 22 and S.D. = 4. Find the mean.

10. The Table shows the distribution of age at the time of first delivery of 65 cows. Find mean deviation from mean and median.

<table>
<thead>
<tr>
<th>Age</th>
<th>18-22</th>
<th>22-26</th>
<th>26-30</th>
<th>30-34</th>
<th>34-38</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>20</td>
<td>30</td>
<td>11</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

11. Read the data given below and find the combined mean, S.D. and coefficient of variation.

\[ n_1 = 15 \quad n_2 = 20 \]
\[ X_1 = 40 \quad X_2 = 50 \]
\[ \sigma_1 = 3 \quad \sigma_2 = 5 \]
12. Mean and standard deviation of lengths of tails of 8 rats were found to be 4.7 cm and 0.8 cm respectively. However, one reading was taken as 3.6 cm instead of 6.3 cm, find the corrected mean and standard deviation.

Answers to unsolved questions

1. Rs. 85.69
2. 31.64
3. 1:1
4. 34
5. 116.7 cm
6. 123.33
7. Q.D. = 11.07; Coefficient of Q.D. = 0.338
8. 11.5
9. 18.18
10. 2.462
11. Combined mean = 45.7, Combined SD = 6.53, CV in % = 14.29
12. Corrected mean = 5.0375 cm; corrected S.D. = 0.8336 cm.
CHAPTER 3

Probability and Probability Laws

Basic terminology used in probability theory

Experiment

An operation that results in definite outcome is called an experiment. Tossing a coin is an experiment if it shows head (H) or tail (T) on falling. If a coin stands on its edge, then it is not considered as an experiment.

Random experiment

When the outcome of an experiment cannot be predicted, then it is called random experiment or stochastic experiment.

Sample space

Sample space or total number of outcomes of an experiment is the set of all possible outcomes of a random experiment and is denoted by ‘S’.

Example

In tossing of coins, the outcomes are head or tail. The head is denoted as ‘H’ and the tail as ‘T’. In tossing two coins, the sample space ‘S’ is given by

\[ S = \{HH, HT, TH, TT\} \]

If the number of outcome is finite then it is called as finite sample space, otherwise it is called as an infinite sample space.

Event

Event may be a single outcome or combination of outcomes. Event is a subset of sample space.

Example

In tossing a coin getting a head is (event A) a single outcome. Therefore

\[ P(A) = \frac{1}{2} \]
In tossing two fair coins, for getting a head (event A) the possible combinations of outcomes are HT and TH. The sample space consists of HH, HT, TH and TT. Therefore, 
\[ P(A) = \frac{1}{2} \]

**Equally likely events**

Two or more events are said to be equally likely if they have equal chance of occurrence.

Example In tossing an unbiased coin, getting head and tail are equally likely.

**Mutually exclusive events**

Two or more events are said to be mutually exclusive if the occurrence of one prevents the occurrence of other events.

Example In tossing a coin, if head falls, it prevents the occurrence of tail and vice versa.

**Exhaustive set of events**

A set of events is exhaustive if one or other of the events in the set occurs whenever the experiment is conducted. It can be defined also as the set whose sum of sample points forms the total sample point of the experiment.

Complementation of an event

The complement of an event is given by:
\[ P(A^c) = 1 - P(A) \]

**Independent events**

Two events are said to be independent of each other if the occurrence of one is not affected by the occurrence of other or does not affect the occurrence of the other.

Example

Consider tossing of three fair coins as shown in figure below. Then
\[ S = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\} \]

Let,

A be the event of getting three heads
B be the event of getting two heads
C be the event of getting one head
D be the event of not getting a head.

Then, the outcomes for events A, B, C and D are:

A = \{HHH\}; B = \{HHT, HTH, THH\}; C = \{HTT, THT, TTH\}; D = \{TTT\}

Events A, B, C and D are mutually exclusive and exhaustive but not equally likely.
Classical/ Mathematical/ Priori approach

Under this approach the probability of an event is known before conducting the experiment.

The following are some of the examples of classical approach.

a) Getting a head in tossing a coin
b) Drawing a king from well shuffled pack
c) Getting a ‘6’ in throwing a die.

The probability of event ‘A’ is defined as:

\[ P(A) = \frac{m}{n} \]

Where, ‘m’ is the number of favourable outcomes, ‘n’ is the total number of outcomes of the experiments. However, it is not possible to give probability to all events of our life. We cannot attach a definite probability to the event ‘that it will rain today’.

Statistical/ Relative Frequency/ Empirical/ Posteriori approach

Under this approach the probability of an event is arrived at after conducting an experiment. If we want to know the probability that a particular household in an area will have two earning members, then we have to gather data on all household in that area and arrive at the probability. The greater number of households surveyed, the more accurate will be the probability arrived. The probability of an event ‘A’ in this case is defined as:

\[ P(A) = \lim_{n \to \infty} \frac{m}{n} \]

In real life, it is not possible to conduct experiments because of high cost or of destructive type experiments or of vast area to be covered.

Subjective approach

Under this approach the investigator or researcher assigns probability to the events either from his experience or form past records. It is more suitable when the sample size is ten or less than ten. The investigator has full knowledge about the characteristics of each and every individual. However, there is a chance of personal bias being introduced in such probability.

Axiomatic approach

This approach is based on set theory. The probability of an event is defined as:

\[ P(A) = \frac{n(A)}{n(S)} \; ; \; \text{Such that} \]
a. \(0 \leq P(A_i) \leq 1\)  

b. \(\sum P(A_i) = 1\) for \(i = 1\) to \(n\)

Where, \(A_i\) is ‘\(n\)’ mutually exclusive and exhaustive events.

**Theorem of addition of probability - Exhaustive events**

Two or more events are said to be mutually exclusive if they cover between them all possible elementary events in relation to the experiment. The union set of exhaustive events is the set of all elementary events in relation to the experiment i.e. its sample space. Where

- \(S = \{1, 2, 3, 4, \ldots, 99, 100\}\)
- \(E = \{2, 4, 6, 8, \ldots, 98, 100\}\)
- \(O = \{1, 3, 5, 7, \ldots, 97, 99\}\)
- \(F = \{5, 10, \ldots, 95, 100\}\)
- \(T = \{3, 6, 9, 12, \ldots, 96, 99\}\)

Events "E" and "O" together form exhaustive events, since \(E \cup O = S\)

Any events combined with the exhaustive events ("E" and "O" here) would also form exhaustive events,

- "E", "O" and "F" are exhaustive events \(\Rightarrow E \cup O \cup F = S\).
- "E", "O" and "T" are exhaustive events \(\Rightarrow E \cup O \cup T = S\)
- "E", "O", "F" and "T" are exhaustive events \(\Rightarrow E \cup O \cup F \cup T = S\)

Probability of occurrence of the sample space is a certainty i.e. its probability is 1.

\[
P(S) = \frac{n(S)}{n(S)} = 1
\]

Since the union of exhaustive events is equal to the sample space, the probability of occurrence of the event representing the union of exhaustive events is a certainty i.e. its probability is 1.

Where \(E \cup O = S\), \(P(E \cup O) = P(S) \Rightarrow P(E \cup O) = 1\)
Theorem of Multiplication of Probability:

The probability of the simultaneous occurrence of two events $E_1$ and $E_2$ is equal to the product of the probability of $E_1$ and the conditional probability of $E_2$ given that $E_1$ has occurred or it is equal to the product of the probability of $E_2$ and conditional probability of $E_1$ given $E_2$ (using multiplication theorem).

$$P(E_1 \cap E_2) = P(E_1).P(E_2/E_1), \text{ whenever } P(E_1) \neq 0$$

$$= P(E_2).P(E_1/E_2), \text{ whenever } P(E_2) \neq 0$$

If $E_1$ and $E_2$ are independent, then $P(E_1/E_2) = P(E_2)$ so that

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

Example

A coin is tossed 3 times. What is the application of the multiplication theorem.

Solution:

Probability of head in the first toss $P(A) = \frac{1}{2}$

Probability of head on the second toss $P(B) = \frac{1}{2}$

Probability of head on the third toss $P(C) = \frac{1}{2}$

Since the events are independent, the probability of getting all heads in three tosses is:

$$P(ABC) = P(A) \times P(B) \times P(C) \quad \text{(using multiplication theorem)}$$

$$P(ABC) = \left(\frac{1}{2}\right)^3 = \frac{1}{8}$$

Theorem of total probability:

If $A$ and $B$ be any two events in a sample space $S$, then the probability of occurrence of at least one of the events $A$ and $B$ is given by

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

Proof: From set theory, we know that
\[ n(A \cup B) = n(A) + n(B) - n(A \cap B) \]

Dividing both sides by \( n(S) \), we get

\[ n(A \cup B) / n(S) = n(A) / n(S) + n(B) / n(S) - n(A \cap B) / n(S) \]

Or

\[ P(A \cup B) = P(A) + P(B) - P(A \cap B) \]

\[ \frac{A}{B} \]

\[ A \cap B \]

\[ \frac{A \cup \bar{B}}{S} \]

**Note:**

i. If \( A \) and \( B \) are mutually exclusive, events, then \( A \cap B = 0 \) and hence

\[ P(A \cap B) = 0 \]

\[ P(A \cup B) = P(A) + P(B) \]

ii. Two events \( A \) and \( B \) are mutually exclusive if and only if

\[ P(A \cup B) = P(A) + P(B) \]

iii. \( 1 = P(S) = P(A \cup A') = P(A) + P(A') \) 

\[ [A \cap A' = 1] \]

Or

\[ P(A') = 1 - P(A) \]

**Bayes Theorem of Probability:**

If \( E_1, E_2, \ldots, E_n \) are \( n \) non-empty events which constitute a partition of sample space \( S \), i.e. \( E_1, E_2, \ldots, E_n \) are pair wise disjoint and \( E_1 \cup E_2 \cup \ldots \cup E_n = S \) and \( A \) is any event of non-zero probability, then

\[ P(E_i/A) = \frac{P(E_i)P(A/E_i)}{\sum_j P(E_j)P(A/E_j)} \text{ for any } i = 1,2,3,\ldots,n \text{ and } j=1,2,3,\ldots,n. \]

The events \( E_1, E_2, \ldots, E_n \) are called hypothesis.

The probability \( P(E_i) \) is called the prior probability if the hypothesis of \( E_i \).

The conditional probability \( P(E_i/A) \) is called the a posteriori probability of the hypothesis \( E_i \).
Bayes theorem is also called the formula for the probability of “causes”. Since the E_i’s are a partition of the sample space S, one and only one of the events E_i occurs (i.e. one of the events E_i must occur and only one can occur).

**Formula for Bayes Theorem of Probability:**

The formula, \( P(E_i/A) = \frac{P(E_i)P(A/E_i)}{\sum P(E_j)P(A/E_j)} \) gives us the probability of a partition E_i (i.e. a “Cause”), given that the event A has occurred.

Example:

The urns contain 6 green, 4 black; 4 green, 6 black and 5 green, 5 black balls respectively. Randomly selected an urn and a ball is drawn from it. If the ball drawn is Green, find the probability that it is drawn from the first urn.

Solution:

Let E_1, E_2, E_3 and A be the events defined as follows:

\( E_1 = \text{urn first is chosen, } E_2 = \text{urn second is chosen,} \)

\( E_3 = \text{urn third is chosen, and } A = \text{ball drawn is Green.} \)

Since there are three urns and one of the three urns is chosen at random, therefore

\( P(E_1) = P(E_2) = P(E_3) = \frac{1}{3} \)

If E_1 is already occurred, then urn first has been chosen which contains 6 Green and 4 Black balls. The probability of drawing a green ball from it is 6/10.

So, \( P(A/E_1) = \frac{6}{10} \)

Similarly, we have \( P(A/E_2) = \frac{4}{10} \) and \( P(A/E_3) = \frac{5}{10} \)

We are required to find \( P(E/A) \), i.e. given that the ball drawn is Green, what is the probability that it is drawn from the first urn.

By Bayes’ theorem, we have

\[
P(E_1/A) = \frac{P(E_1)P(A/E_1)}{P(E_1)P(A/E_1) + P(E_2)P(A/E_2) + P(E_3)P(A/E_3)}
\]

\[
= \frac{(1/3)*(6/10)}{(1/3)*(6/10) + (1/3)*(4/10) + (1/3)*(5/10)}
\]

\[
= \frac{6}{15} = \frac{2}{5}
\]
Conditional Probability

When we are solving Conditional Probability Examples, we deal with two events, say A and B, sometimes these events are so related to each other, that the probability will depend on whether the other event has already occurred. Conditional probability can be defined as the probability of an event given that another event has already occurred.

Let A and B be any two events associated with a random experiment. The probability of occurrence of event A when the event B has already occurred is called the conditional probability of A when B is given and is denoted as P(A/B). The conditional probability P(A/B) is meaningful.

Solution: By definition,

\[ P(A/B) = \text{Probability of occurrence of event A when the event B as already occurred}. \]

\[ = \frac{\text{Number of cases favourable to B which are also favourable to A}}{\text{Number of cases favourable to B}} \]

\[ \therefore P(A/B) = \frac{\text{Number of cases favourable to } A \cap B}{\text{Number of cases favourable to B}} \]

Also,

\[ P(A/B) = \frac{\text{Number of cases favourable to } A \cap B}{\text{Number of cases in the sample space}} \times \frac{\text{Number of cases in the sample space}}{\text{Number of cases favourable to B}} \]

\[ \therefore P(A/B) = \frac{P(A \cap B)}{P(B)} \quad \text{...(1)} \]

Formula for conditional probability

If P(A) \( \neq 0 \), then

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

If A and B are mutually exclusive events, then
If \( A \) and \( B \) are mutually exclusive events, then \( A/B \) and \( B/A \) are impossible events.

**Example:** A pair of dice is thrown. If it is known that one die shows a 4, what is the probability that

a) The other die shows a 5

b) The total of both the die is greater than 7

**Solution:**

Let \( A \) be the event that one die shows up 4. Then the outcomes which are favourable to \( A \) are

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

\( \Rightarrow n(A) = 11 \)

(a) Let \( B \) be the event of getting a 5 in one of the dies. Then the outcomes which are favourable to both \( A \) and \( B \) are \( (4, 5), (5, 4) \)

\( \Rightarrow n(A \cap B) = 2 \)

\[
\therefore P(B/A) = \frac{n(A \cap B)}{n(A)} = \frac{2}{11}
\]

(b) Let \( C \) be the event of getting a total of both the die greater than 7.

The outcomes which are favourable to both \( C \) and \( A \).

\( (4, 4), (4, 5), (4, 6), (5, 4), (6, 4) \)

\( n(C) = 5 \)
Note that in the above example P(B) and P(B/A) are different.

\[ P(B/A) = \frac{2}{11} \text{ where as } P(B) = \frac{11}{36}. \]

Similarly P(C) and P(C/A) are different.

**Solved problem:**

The probabilities that Mr. Aravind, Mr. Anand and Mr. Akil will become vice-president of a company are 0.40, 0.35 and 0.25, respectively. The probabilities that they will introduce new product are 0.10, 0.15 and 0.20 respectively. What is the probability that Mr. Anand introduced a new product by becoming vice-president?

Solution: Let us assume the following:

Let 'A_1' be the event that Mr. Aravind became vice-president
Let 'A_2' be the event that Mr. Anand became vice-president
Let 'A_3' be the event that Mr. Akil became vice-president
Let 'B' be the event that a new product was introduced.

We are given that

\[ P(A_1) = 0.4, \quad P(A_2) = 0.35, \quad P(A_3) = 0.25. \]

\[ P\left(\frac{B}{A_1}\right) = 0.10, \quad P\left(\frac{B}{A_2}\right) = 0.15, \quad P\left(\frac{B}{A_3}\right) = 0.20 \]

\[ P(A \cap B) = P(A) \cdot P\left(\frac{B}{A_1}\right) \]

\[ and \quad P(B) = \sum P(A_1 \cap B) \]

\[ = \sum P\left(\frac{B}{A_1}\right) \]

\[ P(A_2 | B) = \frac{P(B | A_2) \cdot P(A_2)}{P(B)} = \frac{[P(B | A_2) \cdot P(A_2)]}{\sum P(A_1 \cap B)} \]

The required probability are calculated and represented in the Table below.
| Event $A_i$ | Prior Probability $P(A_i)$ | Conditional probability $P(B/A_i)$ | Joint probability $P(A_i \cap B)$ | Posterior probability $P(A_i | B)$ |
|-------------|--------------------------|-----------------------------|-----------------------------|-------------------------------|
| $A_1$       | 0.40                     | 0.10                        | 0.0400                      | $\frac{0.0400}{0.1425} = 0.2807$ |
| $A_2$       | 0.35                     | 0.15                        | 0.0525                      | $\frac{0.0525}{0.1425} = 0.3684$ |
| $A_3$       | 0.25                     | 0.20                        | 0.0500                      | $\frac{0.0500}{0.1425} = 0.3509$ |
| Total       | 1.00                     | $P(B) = 0.1425$             |                             | $1.0000$                      |

Therefore, the required probability $P(A_2 | B) = 0.3684$

**Self assessment questions**

1. To which approach the following probability estimates belong to:
   i. The probability that India will win the game.
   ii. The probability that Mr. Ram will resign from the post.
   iii. Probability of drawing a red card.
   iv. Probability that you will go to America this year.

2. Find the probability in the following cases:
   i. Getting an even number when a die is thrown.
   ii. Selecting two ‘y’s’ from the letters, x, x, x, x, y, y, y.
   iii. Selecting a King and Queen from a pack of cards, when two cards are drawn at a time.
   iv. Getting 53 Mondays in ordinary year.

3. Given $P(A) = 0.6$, $P(B) = 0.7$ and $P(A \cap B) = 0.5$. Find $P (A \cup B)$?

4. State whether the following questions are true or false-
   i. Bayers’ probability estimates sample value
   ii. Conditional probability can incorporate costs
   iii. Bayers’ probability gives up to date information.

**Unsolved questions**
1. Define independent events.

2. The probability of Mr. Sunil solving the problem is $\frac{3}{4}$. The probability of Mr. Anish to solving is $\frac{1}{4}$. What is the probability that a given problem will be solved?

3. The probability that a contractor will get an electrical job is 0.8, he will get a plumbing job is 0.6 and he will get both 0.48. What is the probability that he get at least one? Is the probabilities of getting electrical and plumbing job are independent?

4. A box contains 4 red and 5 blue similar rings. What is the probability of selecting at random two rings: i. Having same colour, ii. Having different colour.

5. If $P(A \cap B) = \frac{1}{2}$ and $P(B) = \frac{2}{3}$, find $P(A/B)$?

6. If $P(A \cup B) = 0.8$, $P(A) = 0.6$, $P(B) = 0.7$, find $P(B/A)$?

7. A shopkeeper sells two types of articles, namely washing machine and ovens. He had 10 similar washing machines and 20 similar ovens. 2 of the washing machine and 5 of the oven have 50% discount on them. A customer selects an article. What is the probability it is a washing machine or an article with discounted price.

8. The probability that a bomb hitting a target is $\frac{2}{5}$. If four bombs are dropped on a bridge, what is the probability that it will be destroyed?

9. The probability that a company A will survive for 20 years is 0.6. The probability that its sister concern will survive for 20 years is 0.8. What is the probability that at least one of them will survive for 20 years?

10. A recently developed car has two important components A and B. The probability of failure of A and B are 0.2 and 0.1. What is the probability that the car will fail?

11. The probability that a football player will play an ordinary ground is 0.5 and on green turf is 0.4. The probability will get knee injury when playing an ordinary ground is 0.07 and that a green turf is 0.04. What is the probability that he got knee-injury due to the play on ordinary ground?

12. Find the $E(X)$ and $Var(X)$ for the distribution of a random variable, X represented in table below:
### Solutions to Self Assessment Questions

1. **i.** Relative frequency  
   **ii.** Subjective  
   **iii.** Classical  
   **iv.** Subjective

2. **i.** $\frac{1}{2}$  
   **ii.** $\frac{1}{7}$  
   **iii.** $\frac{8}{663}$  
   **iv.** $\frac{1}{7}$

3. 0.8

4. **i.** False  
   **ii.** False  
   **iii.** True

5. **i.** 1  
   **ii.** Mean  
   **iii.** $[E(X)]^2$

### Solutions to Unsolved Questions

1. Refer text material
2. 13/6
3. 0.92, yes
4. **i.** $\frac{4}{9}$  
   **ii.** $\frac{5}{9}$
5. $\frac{3}{4}$
6. $\frac{5}{8}$
7. 2/3
8. $\frac{544}{625}$
9. 0.92
10. 0.28
11. 21/29
12. $E(X) = \frac{7}{3}, Var(X) = \frac{115}{18}$
CHAPTER 4

Random Variable and Mathematical Expectation

It has been a general notion that if an experiment is conducted under identical conditions, values so obtained would be similar. But the experiences of people have dispelled this belief. Observations are always taken about a factor or character under study which can take different values, and the factor or character is termed as variable. A rule that assigns a real number to each outcome (sample points) is called random variable.

Random variable (r.v.)

A random variable $X$ is a real valued function, $X(x)$, of the elements of the sample space $\Omega$ where $x$ is an element of the sample space. It should be noted that range of random variable will be a set of real numbers.

Illustration: If we measure the height of people, the random variable,

$$X(x) = \{ x : x \text{ is any real positive number} \}$$

Random variables are of two types (i) Discrete random variable and (ii) Continuous random variable.

Discrete Random variable:

A random variable $X$, which can take only a finite number of values in an interval of the domain, is called discrete random variable. A few examples are: (i) If we throw two dice at a time and note the sum of spots which turn up on the upper face, the discrete r.v.,

$$X(x) = \{ x : x = (x_1 + x_2) = 2, 3, \ldots, 12 \}$$

Where, $x_1$ is the number of spots on the upper face of one die and $x_2$ is the number of spots on the other die. (ii) The random variable denoting the number of students in a class is,

$$X(x) = \{ x : x \text{ is any positive integer} \}$$

Continuous random variable:

A random variable $X$, which can take any value in the domain, or when its range ‘$R$’ is an interval or the union of intervals on the real line, is called a continuous random variable.

Note that the probability of any single $x$, a value of $X$, is zero i.e.

$$P (X=x) = 0$$
Example: The height of students in India lies between 3 and 6 feet. The continuous r.v.
\[ X(x) = \{ x: 3 \leq x \leq 6 \} \]

Example. The maximum life of electric bulbs is 2000 hours. The continuous r.v.,
\[ X(x) = \{ x: 0 \leq x \leq 2000 \} \]

**Distribution function**

The distribution of a random variable \( X \) is a function \( F_x(x) \) for a real value \( x \) which is the probability of the event \( (X \leq x) \), i.e.
\[ F_x(X) = P(X \leq x) \]

It is evident from the expression that the distribution function is the probability that \( X \) takes a value in the interval \((-\infty, x)\).

**Discrete Probability Distribution:** If a random variable is discrete. Its distribution will also be discrete, except in some exceptional situations. For a discrete random variable \( X \), the distribution function or cumulative distribution is given by \( P(x) \) and is same as given by (6.1) i.e.
\[ P(x) = P(X \leq x) \]

**Discrete Probability Function:** The probability function is the probability of a discrete random variable \( X \), which takes the value \( x \) and is denoted by \( p(x) \) i.e.
\[ P(x) = P(X = x) \]

The probability function always possesses the following properties:

i. \( P(x) > 0 \) for all \( x \) in sample space \( \Omega \).
ii. \( \sum_{all \ x} p(x) = 1 \).
Mathematical expectation

The expected value of a random variable $X$ is given as

$$E(X) = \sum_{\text{all } x} x \cdot p(x)$$

$E(X)$ is also known as theoretical average value. In general, the expected value of a function $H(X)$ of a discrete random variable $X$ is given as:

$$E\{H(X)\} = \sum_{\text{all } x} H(x) \cdot p(x)$$

Some Results on Expectation:

If $X$ and $Y$ are two variates defined over the same sample space and if $E(X)$ and $E(Y)$ exist, then

$$E(X + Y) = E(X) + E(Y)$$

Also, if $X$ and $Y$ are independent, then

$$E(XY) = E(X) \cdot E(Y)$$

Consider a variable $X$ and constant $C$. The expected values of $X$ in relation to $C$ are as follows:

$$E(C) = C$$

$$E(CX) = CE(X)$$

$$E(X + C) = E(X) + C$$

If $X$ and $Y$ are two variables and $C_1$ and $C_2$ are any constants, then

$$E(C_1X + C_2Y) = C_1E(X) + C_2E(Y)$$

If $X \leq Y$, then

$$E(X) \leq E(Y)$$

The operations given form (6.8) to (6.12) are also true for functions of $X$ & $Y$. 
Moments

The \( r \)\textsuperscript{th} moment of a random variable \( X \) is defined as the expected value of \( X^r \) where \( r \) is an integer i.e. \( r = 1, 2, \ldots \). The \( r \)\textsuperscript{th} moment \( E(X^r) \), about the origin (zero), is known as the \( r \)\textsuperscript{th} raw moment and is generally denoted by \( \mu'_r \). Thus

\[
\mu'_r = E(X^r)
\]

Similarly, the \( r \)\textsuperscript{th} moment of a random variable \( X \) about the mean of a frequency distribution, is given by the formula.

\[
\mu_r = E(X - \mu)^r
\]

Where, \( \mu \) is symbolically known as the \( r \)\textsuperscript{th} central moment of a random variable \( X \).

Moment Generating function (M.G.F.)

From the discussion of moments, it is apparent that moments play an important role in the characterization of various distributions. Hence, to know the distribution, we need to find out the moments. For this the moment generating function is a final device. The moment generating function is a special form of mathematical expectation, and is very useful in deriving the moments of a probability distribution.

**Definition:** If \( X \) is a random variable, then the expected value of \( e^{tx} \) is known as the moment generating function, provided the expected value exists for every value of \( t \) in an interval, \(-h < t < h\) where, \( h \) is some positive real value. The moment generating function which is denoted as \( m_x(t) \) for a discrete random variable is:

\[
m_x(t) = E(e^{tx}) = \sum_{x} e^{tx} p_x(x)
\]

\[
= \sum_{x} \left(1 + tx + \frac{t^2 x^2}{2!} + \frac{t^3 x^3}{3!} + \ldots\right) p_x(x)
\]

\[
= 1 + t \mu'_1 + \frac{t^2}{2!} \mu'_2 + \frac{t^3}{3!} \mu'_3 + \ldots = \sum_{r=0}^{\infty} \frac{t^r}{r!} \mu'_r
\]

Characteristic Function (C.F.)

The moment generating function does not exist for every distribution. Hence, another function, which always exists for all the distribution, is known as characteristic function. It is the expected value of \( e^{itx} \), where \( i = \sqrt{-1} \) and \( t \) is a real valued continuous variable. Let the characteristic function of a random variable \( X \) be denoted by \( \phi_x(t) \), then

\[
\phi_x(t) = E(e^{itx})
\]
For a discrete variable $X$ having the probability function $p(x)$, the characteristic function.

$$\phi_x(t) = \sum_{\text{all } x} e^{itx} p(x)$$

And for a continuous variable $X$ having density function $f(x)$, such that $a < X < b$, the characteristic function,

$$\phi_x(t) = \int_a^b e^{itx} f(x) \, dx$$

It is easy to prove that $\phi_x(0) = 1$.

**Uniqueness theorem:** Two distribution functions are identical if their characteristic functions are identical.

**Inversion theorem:** Let $F$ be a distribution function and $\phi_x(t)$ be its characteristic function. If $a$ and $b$ are two points on a real line such that $a < b$, then we have,

$$F(b) - F(a) = \lim_{r \to \infty} \frac{1}{2\pi} \int_r^r e^{-i\alpha x} - e^{-i\beta x} \phi_x(t) \, dt$$
CHAPTER- 5

**Probability Distributions**

As the random variables are discrete and continuous, the probabilities associated with random variables are also discrete and continuous. The listing of all the probable outcomes in a random experiment along with their respective probabilities is called the probability distribution.

**Discrete probability distributions**

A discrete probability distribution consists of all possible values of a discrete random variable along with their corresponding probabilities, binomial, Bernoulli, Poisson are all examples of discrete probability distributions.

**Continuous probability distribution**

In a continuous probability distribution the variable under consideration assumes any value within a given range. Hence, it is very difficult to list all values. One example of continuous probability distribution is the distribution of normal variable.

1. **Bernoulli distributions**

A variable which assumes values 1 and 0 with probability ‘p’ and ‘q’ (where, q = 1-p) is called Bernoulli variable. It has only one parameter ‘p’ for different values of ‘p’ (0≤p≤1), we get different Bernoulli distributions. In these distributions, ‘1’ represents the occurrence of success and ‘0’ represents the occurrence of failure.

In other words the assumption for the distribution is outcome of an experiment. It is of dichotomous nature, that is, success/ failure, present/ absent, defective/ non defective, yes/ no and so on.

**Example**

When a fair coin is tossed as shown in figure below, the outcome is either head or tail. The variable ‘X’ assumes ‘1’ or ‘0’

**Solved problem** : An unbiased coin is tossed six times. What is the probability that the tosses will result in:

i. Exactly two heads
ii. At least five heads
iii. At most two heads
iv. Not greater than one head
v. Not less than five heads
vi. At least one head
**Solution:** let ‘A’ be the event of getting head. Given that

\[ P = \frac{1}{2}, \quad q = \frac{1}{2}, \quad n=6 \]

\[ \therefore \text{ Binomial distribution is } (\frac{1}{2} + \frac{1}{2})^6 \]

i. The probability that the tosses will result in exactly two heads is given by:

\[ P(X = 2) = \binom{6}{2} \left( \frac{1}{2} \right)^2 \left( \frac{1}{2} \right)^4 = \frac{15}{64} \]

Therefore, the probability that the tosses will result in exactly two heads is 15/64.

ii. The probability that the tosses will result in at least five heads is given by

\[ P(X \geq 5) = P(X=5) + P(X=6) = \binom{6}{5} \left( \frac{1}{2} \right)^5 \left( \frac{1}{2} \right)^1 + \binom{6}{6} \left( \frac{1}{2} \right)^6 \]

\[ = \frac{6 \times 5}{2^6} \times \frac{1}{2} + \frac{1 \times 1}{2^6} = \frac{7}{64} \]

Therefore, the probability that the tosses will result in at least five heads is 7/64.

iii. The probability that the tosses will result in at most two heads is given by:

\[ P(X \leq 2) = P(X=0) + P(X=1) + P(X=2) \]

\[ = \left( \frac{1}{2} \right)^6 + \binom{6}{1} \left( \frac{1}{2} \right)^1 \left( \frac{1}{2} \right)^5 + \binom{6}{2} \left( \frac{1}{2} \right)^2 \left( \frac{1}{2} \right)^4 \]

\[ = \frac{1}{64} + \frac{6 \times 1}{2^6} \times \frac{1}{2} + \frac{1 \times 1 \times 15}{2^6} = \frac{22}{64} = \frac{11}{32} \]

Therefore, the probability that the tosses will result in at most two heads is 11/32.

iv. The probability that the tosses will result in not greater than one head is given by:

\[ P(X \leq 1) = P(X = 0) + P(X = 1) = \frac{1}{64} + \frac{6}{64} = \frac{7}{64} \]

Therefore, the probability that the tosses will result in not greater than one head is 7/64.

v. The probability that the tosses will result in not less than five heads is given by:

\[ P(X \leq 5) = P(X = 5) + P(X = 6) = \frac{6}{2^6} + \frac{1}{2^6} = \frac{7}{64} \]

Therefore, the probability that the tosses will result in not less than five heads is 7/64.

vi. The probability that the tosses will result in at least one head is given by:

\[ P(X \leq 1) = 1 - P(X = 0) = 1 - \frac{1}{2^6} = 1 - \frac{1}{64} = \frac{63}{64} \]
Therefore, the probability that the tosses will result in at least one head for different values of ‘x’.

### 2. Binomial distribution

Binomial distribution was discovered by James Bernoulli (1654-1705) in the year 1700. A random variable X is said to follow binomial distribution if it assumes only non-negative values.

A **binomial random variable** is the number of successes \( x \) in \( n \) repeated trials of a binomial experiment. and its probability mass function is given by:

\[
P(x) = \binom{n}{x} p^x q^{n-x}
\]

where, \( x = 0, 1, 2, 3, \ldots, n \)

If \( x \leq 0 = 0, 1, 2, 3, \ldots, n \) it is not binomial distribution.

The two independent constants \( n \) and \( p \) in the distribution are known as the parameters of the distribution. ‘\( n \)’ is also sometimes, known as the degree of the binomial distribution.

Binomial distribution is a discrete distribution as \( X \) can take only the integral value, viz., 0, 1, 2, \ldots, \( n \). any random variable which can follow binomial distribution is known as binomial variate.

- \( x \): The number of successes that result from the binomial experiment.
- \( n \): The number of trials in the binomial experiment.
- \( P \): The probability of success on an individual trial.
- \( Q \): The probability of failure on an individual trial. (This is equal to \( 1 - P \).)

\( b(x; n, P) \): Binomial probability - the probability that an \( n \)-trial binomial experiment results in exactly \( x \) successes, when the probability of success on an individual trial is \( P \).

\( ^nC_x \): The number of combinations of \( n \) things.

**Condition which follow binomial distribution:**

A **binomial experiment** is a probability experiment that satisfies the following four requirements:

1. Each trial can have only two outcomes or outcomes that can be reduced to two outcomes. These outcomes can be considered as either success or failure.
2. There must be a fixed ‘\( n \)’ number of trials.
3. The outcomes of each trial must be independent of each other.
4. The probability of a success must remain the same for each trial.
Properties of Binomial distribution :-

i. The shape and location of binomial distribution change as p change for a given ‘n’

ii. The mode of B.D.is equal to the value of x which has largest probability.

iii. The mean & mode are equal if np (mean) is an integer, eg.- n=6 & p=0.5 then then mean=mode=3.

iv. For fixed n, both mean & mode $\uparrow$ as p $\uparrow$.

v. If P=constant, & n $\uparrow$, than B.D. move to right flattens & spread out.

Constants of Binomial distribution :-

- The mean of the distribution ($\mu_x$) is equal to $nP$.
- The variance ($\sigma^2_x$) is $nPq$.
- Variance is always less than mean.
- The standard deviation ($\sigma_x$) is $\sqrt{npq}$
- The two independent constant n and p is the distribution are known as parameter of distribution

Example : Ten coins are thrown simultaneously. Find the probability of getting at least seven heads.

Solution:- P= probability of getting a head $= \frac{1}{2}$

q=probability of not getting a tail $= \frac{1}{2}$

The probability of getting x head in a random throw of 10 coins is:
\[
\Rightarrow P(x) = \binom{n}{x} p^x q^{n-x}
\]
\[
\Rightarrow 10 C_x \left( \frac{1}{2} \right)^x \left( \frac{1}{2} \right)^{10-x}
\]
\[
\Rightarrow 10 C_x \left( \frac{1}{2} \right)^{x+10-x}
\]
\[
\Rightarrow 10 C_x \left( \frac{1}{2} \right)^{10}
\]

Now, \((P \geq 7)\)
\[
\Rightarrow P(7) + P(8) + P(9) + P(10)
\]
\[
\Rightarrow 10 C_7 \left( \frac{1}{2} \right)^{10} + 10 C_8 \left( \frac{1}{2} \right)^{10} + 10 C_9 \left( \frac{1}{2} \right)^{10} + 10 C_{10} \left( \frac{1}{2} \right)^{10}
\]
\[
\Rightarrow \left( \frac{1}{2} \right)^{2} \left[ 10 C_7 + 10 C_8 + 10 C_9 + 10 C_{10} \right]
\]
\[
\Rightarrow \left( \frac{1}{2} \right)^{2} (120 + 45 + 10 + 1)
\]
\[
\Rightarrow \frac{176}{1024}
\]

**Example**: A and B play a game in which their chances of winning are in the ratio 3:2. Find that the chance of winning ‘A’ for at least 3 games out of 5 games.

**Solution**:

\[ P = \frac{3}{5}, \quad q = \frac{2}{5} \]

Total no. of games = 5

A’ will win at least 3 times = \( \frac{3}{5} \)

\therefore B’ will win at least 2 times = \( \frac{2}{5} \)

Probability of x –
\[
\Rightarrow P(X) = \binom{5}{x} p^x q^{5-x}
\]
\[
\Rightarrow P(X) = \binom{5}{x} \left( \frac{3}{5} \right)^x \left( \frac{2}{5} \right)^{5-x} \]
\[ C_3 \left( \frac{3}{5} \right)^3 \left( \frac{2}{3} \right)^{5-3} \]

Probability of winning at least 3 games by ‘A’

\[ P(3) + P(4) + P(5) \]

\[ C_3 \left( \frac{3}{5} \right)^3 \left( \frac{2}{5} \right)^{5-3} + C_4 \left( \frac{3}{5} \right)^4 \left( \frac{2}{5} \right)^{4-4} + C_5 \left( \frac{3}{5} \right)^5 \left( \frac{2}{5} \right)^{5-5} \]

\[ C_3 \frac{3^3}{5^3} \left( \frac{2}{5} \right)^{5} + C_4 \frac{3^4}{5^4} \cdot \frac{2}{5} + C_5 \frac{3^5}{5^5} \]

\[ \frac{3^3}{5^3} \left[ C_3 \cdot \frac{4}{25} + C_4 \left( \frac{3}{5} \cdot \frac{2}{5} \right) + C_5 \cdot \frac{9}{25} \right] \]

\[ \frac{3^3}{5^3} \left[ 10 \times \frac{4}{25} + 5 \times \frac{6}{25} + \frac{9}{25} \right] \]

\[ \frac{3^3}{5^3} \cdot \frac{1}{25} \left[ 40 + 30 + 9 \right] \]

\[ \frac{27}{125} \times \frac{1}{25} \left[ 79 \right] \]

\[ \frac{27 \times 79}{125 \times 25} \]

\[ \frac{2133}{3125} \]

\[ 0.68 \]

Interpretation: Probability of winning at least 3 games by ‘A’ is 0.68

**Binomial probability:**

The probability of a success in a binomial experiment can be computed with the following formula.

**Binomial Probability Formula:**

In a binomial experiment, the probability of exactly X successes in n trials is

\[ P(X) = \frac{n!}{(n-X)!X!} \cdot p^X \cdot q^{n-X} \]
An explanation of why the formula works will be given in the following example.

Example: A coin is tossed three times. Find the probability of getting exactly two heads.

Solution: This problem can be solved by looking at the sample space. There are three ways to get two heads.

HHH, HHT, HTH, THH, TTH, THT, HTT, TTT

The answer is $\frac{3}{8}$ or 0.375.

Looking at the problem in the previous example from the standpoint of a binomial experiment, one can show that it meets the four requirements.

1. There are only two outcomes for each trial, heads or tails.
2. There is a fixed number of trials (three).
3. The outcomes are independent of each other (the outcome of one toss in no way affects the outcome of another toss).
4. The probability of a success (heads) is 1/2 in each case.

In this case, $n = 3$, $X = 2$, $p = 1/2$, and $q = 1/2$. Hence, substituting in the formula gives

$$P(2 \text{ Heads}) = \frac{3!}{(3 - 2)!2!} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^1 = \frac{3}{8} = 0.375$$

3. Poisson Distribution

Poisson distribution was discovered by the French mathematician Semeon Denis Poisson (1781-1840).

Poisson distribution is a discrete probability distribution that expresses the probability of a number of event occurring in a fixed period of time if these event occur with a known average rate and independently of the time since the last event.

A random variable $x$ is said to follow a poission distribution if it assumes only non-negative value and its probability mass function is given by :-

$$P(x, \lambda) = P(x = x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad x = 1, 2, 3, \ldots \infty$$
Where \( e = \text{base of the natural logarithm (} e = 2.71828...) \)

\[ x = \text{number of occurrences of an event - the probability of which is given by the function} \]

\[ x! = \text{factorial of } n \]

\( \lambda = \text{a positive real number, equal to the expected number of occurrences that occur during the given interval. For instance, if the events occur on average 4 times per minute, and you are interested in the number of events occurring in a 10 minute interval, you would use as your model a Poisson distribution with } \lambda = 10 \times 4 = 40. \]

When it takes all possible value then -

\[ \Rightarrow \sum_{x=0}^{\infty} P(X = x) = \sum_{x=0}^{\infty} e^{-\lambda} \frac{\lambda^x}{x!} = 1 \]

\[ \Rightarrow e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \]

\[ \Rightarrow e^{-\lambda} \left[ \frac{\lambda^0}{0!} + \frac{\lambda^1}{1!} + \frac{\lambda^2}{2!} + \ldots \right] \quad 0 = \text{otherwise} \]

\[ \Rightarrow e^{-\lambda} \left[ 1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \ldots \right] \]

\[ \Rightarrow e^{-\lambda} e^{-\lambda} = e^0 = 1 \]

**Poisson distribution is a limiting case of the binomial under the following condition:**

a. \( n, \text{ the number of trials is indefinitely large ,i.e.,} n \to \infty. \)

b. \( P, \text{ The constant probability of success for each trial is indefinitely small, i.e.,} p \to 0. \)

c. \( np = \lambda, \text{is finite} \)

thus \( p = \frac{\lambda}{n}, \quad q = \frac{1-\lambda}{n} \quad \text{where } \lambda \text{ is a positive real number.} \)

**Properties:**

Probability mass function= \( p(x, \lambda) = p(x = x) = \frac{e^{-\lambda} \lambda^x}{x!} \)

Moments of poisson distribution:

- **About origin** -
\[ \Rightarrow \mu_1^1 = (mean) = E(x) = \lambda \]
\[ \Rightarrow \mu_2^1 = E(x(x-1)+x) = \lambda^2 + \lambda \]
\[ \Rightarrow \mu_3^1 = \lambda^3 + 3\lambda^2 + 7\lambda^2 + \lambda \]
\[ \Rightarrow \mu_4^1 = \lambda + 6\lambda^3 + 7\lambda^2 + \lambda \]

- **About mean -**
  \[ \Rightarrow \mu_2 = \lambda (variance) \]
  \[ \Rightarrow \mu_2^2 = (\mu_1^1)^2 \]
  \[ \Rightarrow \lambda^2 + \lambda - (\lambda)^2 \]
  \[ \Rightarrow \lambda^2 + \lambda - \lambda^2 \]
  \[ \Rightarrow \lambda \]
  \[ \Rightarrow \mu_3 = \mu_3^1 - 3\mu_1^1\mu_1^1 + 2(\mu_1^1)^3 \]
  \[ \Rightarrow \lambda^3 + 3\lambda^2 + \lambda - 3\lambda^2 + \lambda \cdot \lambda + 2(\lambda)^3 \]
  \[ \Rightarrow \lambda^3 + \lambda - \lambda(\lambda^2 + \lambda^2) + 2(\lambda)^3 \]
  \[ \Rightarrow \lambda^3 + \lambda^2 - \lambda^2 + 2(\lambda)^3 \]
  \[ \Rightarrow \lambda^3 - 2(\lambda)^3 \]
  \[ \Rightarrow \mu_3 = \lambda \]
  \[ \Rightarrow \mu_4 = \mu_4^1 - \mu_3^1\mu_1^1 + \mu_1^1\mu_1^1 + \mu_1^1 \]
  \[ \Rightarrow \mu_4 = 3\lambda^2 + \lambda \]

**Skewness of poisson distribution :**

\[ \Rightarrow \beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{\lambda^2}{\lambda^3} = \frac{1}{\lambda} \]
Hence Poisson distribution is always a skewed distribution

**Kurtosis of poisson distribution:**

\[ \beta_2 = \frac{\mu_4^2}{\mu_2^3} = \frac{3\lambda^2 + \lambda}{\lambda^3} = \frac{3\lambda^2}{\lambda^3} + \frac{\lambda}{\lambda^3} = 3 + \frac{1}{\lambda} \]
Hence the curve of Poisson distribution is leptokurtic curve.

**A Poisson experiment is a statistical experiment that has the following properties:**

- The experiment results in outcomes that can be classified as successes or failures.
- The average number of successes (\( \mu \)) that occurs in a specified region is known.
- The probability that a success will occur is proportional to the size of the region.
The probability that a success will occur in an extremely small region is virtually zero.

Note that the specified region could take many forms. For instance, it could be a length, an area, a volume, a period of time, etc.

**Example**: A manufacture of cotter pins known that 5% of his product is defective. If he sells cotter pins in boxes of 100 and guarantees that not more than 10 pins will be defective, what is the approximate probability that a box will fail to meet the guaranteed quality?

**Solution**: we are given: \( n=100 \)

Probability of defective pin \( (p) = 5\% = 0.05 \)

Mean number of defective pins \( \lambda = np=100 \times 0.05 =5 \)

The probability of \( x \) defective pin in a box:

\[
P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}
\]

\[
\Rightarrow \frac{e^{-5} 5^x}{x!}; x = 0, 1, 2, \ldots
\]

Probability that a box will fail to meet the guaranteed quality is:

\[
P(X \geq 10) = 1 - P(X \leq 10)
\]

\[
\Rightarrow 1 - \sum_{x=0}^{10} \frac{e^{-5} 5^x}{x!} = 1 - e^{-5} \sum_{x=0}^{10} \frac{5^x}{x!}
\]

\[
\Rightarrow 1 - e^{-5} \left[ \frac{5^0}{0!} + \frac{5^1}{1!} + \ldots + \frac{5^{10}}{10!} \right]
\]

**Example**: 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 a Poisson distribution with mean 1.5. Calculate the proportion of days on which (i) neither car is used, and (ii) the proportion of days on which some demand is refused.

**Solution**: Here the random variable \( X \) has mean \( \lambda = 1.5 \)

The proportion of days on which there are \( x \) demand for a car is given by:

\[
P(X = 0) = \frac{e^{-1.5} (1.5)^x}{x!}; x = 0, 1, 2, \ldots
\]
(i) Proportion of days on which neither car is used is given by:

\[
P(X = 0) = e^{-1.5} = \left[1 - 1.5 + \frac{(1.5)^2}{2!} + \frac{(1.5)^3}{3!} + \frac{(1.5)^4}{4!} - \ldots \right] = 0.2231
\]

(ii) Proportion of days on which some demand is refused is:

\[
P(X \geq 2) = 1 - P(X \leq 2)
\]

\[
= 1 - \left\{ P(X = 0) + P(X = 1) + P(X = 2) \right\}
\]

\[
= 1 - e^{-1.5} \left\{ 1 + 1.5 + \frac{(1.5)^2}{2!} \right\}
\]

\[
= 1 - 0.2231 \times 3.625
\]

\[
= 0.19126.
\]

4. **Negative Binomial Distribution**

A random variable \( X \), the number of failures before the \( r \)th success occurs in a random experiment, which results in either a success or a failure, is called a negative binomial random variable and its probability function with probability ‘\( p \)’ of a success is given by:

\[
P\{nb(x)\} = \binom{x + r - 1}{r - 1} p^r q^x
\]

for , \( i = 0, 1, 2, \ldots \). Where \( r \geq 0, 0 \leq p \leq 1 \) and \( q = 1 - p \)

Probability function can also be written as

\[
P\{nb(x)\} = \binom{-r}{x} p^r (-q)^x
\]

Since \( \binom{x + r - 1}{r - 1} = (-1)^r \binom{-r}{x} \)

The distribution given by \( \binom{-r}{x} \) is also called Pascal’s distribution. The salient feature of this distribution is that the number of successes is fixed and the number of trials is random. It is discrete distribution with mean \( r (1-p) / p \) & variance \( r (1-p)/p^2 \).

5. **Normal Distribution**

In probability theory, the normal (or Gaussian) distribution, is a continuous probability distribution that is often used as a first approximation to describe real-valued random variables that tend to cluster around a single mean value. The graph of the
associated probability density function is “bell”-shaped, and is known as the Gaussian function or bell curve.

The normal distribution is considered the most prominent probability distribution in statistics. There are several reasons for this: First, the normal distribution is very tractable analytically, that is, a large number of results involving this distribution can be derived in explicit form. Second, the normal distribution arises as the outcome of the central limit theorem, which states that under mild conditions the sum of a large number of random variables is distributed approximately normally. Finally, the “bell” shapes of the normal distribution make it a convenient choice for modeling a large variety of random variables encountered in practice.

Normal distribution is the most popular and commonly used distribution. It was discovered by De Moivre in 1733; about 20 years after Bernoulli gave binomial distribution. A random variable X is said to follow normal distribution, if and only if its probability density function (p.d.f.) is

\[
f_X(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}
\]

Where, x is the real value of X, i.e. \(-\infty < X < \infty\).

The variable X is said to be distributed normally with mean \(\mu\) and variance \(\sigma^2\). \(X \sim \mathcal{N}(\mu, \sigma^2)\).

The density function given by above equation has two parameters, namely \(\mu\) and \(\sigma\). Here it can be taken any real value in the range \(-\infty \text{ to } \infty\), where \(\sigma\) is any positive real value, i.e. \(\sigma > 0\).

Since the probability can never by negative

\[F_X(x) \geq 0 \text{ for all } x.\]

In case \(\mu = 0, \sigma = 1\), the density function for X is

\[
f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}
\]

Here, \(-\infty < X < \infty\). Notationally, \(X \sim \mathcal{N}(0, 1)\).

In this situation, the variable X is called the standardized normal variate and distribution given above is called the standardized normal distribution.

**Study of Normal Curve:**

In Normal bell curve distribution, which figure most significantly in statistical theory and in application. *Normal distribution* is also called as the Normal probability
distribution. Let us see how to calculate normal distribution. The normal distribution looks like a bell shaped curve. Hence it is also known as normal bell curve distribution.

The graph of the normal curve is shown above. The shape of the curve is bell. These are the constants which tell how to calculate normal distribution

**Normal distribution as a limiting form of binomial distribution:**

Normal distribution is another form of the binomial distribution under the following condition:

1. n, the number of trial is indefinitely large, i.e. \( n \rightarrow \infty \)
2. neither \( p \) or \( q \) is very small

**Standard Normal Distribution:**

Standard normal Deviate for random variable \( X \sim N(\mu, \sigma^2) \), the location and shape of the normal curve depends on \( \mu \) and \( \sigma \) where \( \mu \) and \( \sigma \) can take any value within their range. Hence, no master table for the area under the curve can be prepared. This difficulty is very well overcome by consideration of a variable \( Z \) where \( Z \sim (X-\mu)/ \sigma \). The variable \( Z \) is always distributed with mean zero and variance unity, i.e. \( Z \sim N(0,1) \)
Properties of normal distribution

1. If we plot a curve for the values of standard normal variate $X$ and the corresponding probabilities, it is of the shape given in above Figure. The curve is bell-shaped and symmetrical about the vertical line at $x=0$. For the random variable $x ~ N (\mu, \sigma^2)$, the curve is symmetrical about the mean. Symmetry implies $f_x(x) = f_x(-x)$.

2. The vertical line at $x = 0$ cuts the curve at its highest point i.e. $f_x(x)$ at $x = 0$ is maximum. The value of $f_x(x)$ decreases as the value of $x$ is other than zero (on the mean $\mu$).

3. The area under the normal curve within the limits – to is unity, i.e.

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

If $\mu = 0$, $\sigma = 1$ then,

$$
\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} = 1
$$

4. On either side of mean $\mu$, the frequency decreases more rapidly within the range $(\mu^2 \pm \sigma^2)$ and gets slower and slower as it goes away from the mean. The frequencies are extremely small beyond the distance of $3\sigma$. As a matter of fact 99.73 per cent units of the population lie within the range $\mu \pm 3\sigma$.

5. Theoretically the curve never touches the X-axis.

6. As the value of $\sigma$ increases, the curve becomes more and more flat and vice-versa.

7. The moment generating function of the general normal distribution is,

$$m_X(t) = e^{(\mu t + \frac{1}{2}\sigma^2 t^2)}$$

8. All central moments of odd order are zero.

9. The characteristic function of the general normal distribution is

$$\phi_X(t) = e^{(\mu t - \frac{1}{2}\sigma^2 t^2)}$$
Normal distribution is most widely used in statistics despite the fact that theoretically a population hardly follows the exact normal distribution. This is due to various reasons which are as follows:

i. Convenience is a strong support for its use in statistics. Extensive tables have been prepared and provided for ordinates and area under the normal curve. This has facilitated the job of the scientist immensely.

ii. If the variable does not follow normal distribution, it can be made to follow after making a suitable transformation like square root, arcsine, logarithm etc.

iii. The most important and convincing reason is that whatever be the original distribution, the distribution of sample mean can, in most of the cases, be approximated to normal distribution, when the sample size is sufficiently increased.

Additionally, every normal curve (regardless of its mean or standard deviation) conforms to the following "rule".

- About 68% of the area under the curve falls within 1 standard deviation of the mean.
- About 95% of the area under the curve falls within 2 standard deviations of the mean.
- About 99.7% of the area under the curve falls within 3 standard deviations of the mean.

**Importance of normal distribution:**

Normal distribution has following important role:

1) Most of the distribution occurring in practice, e.g.- binomial, Poisson, hypergeometric distribution etc. an be approximated by normal distribution, e.g. student’s t, snedecor’s f chi square distribution , etc. tend to normally for large sample.

2) Even if a variable is not normally distributed, it can sometime be brought to normal form by simple transformation of variable.
3) The theory of small sample tests, viz. t,F, test, $x^2$ tests, etc. is based on the fundamental assumption that the parent population from which the samples have been drawn follow normal distribution.

4) If $X \sim N(\mu, \sigma)$, then $p(\mu - 3\sigma \leq x \leq \mu + 3\sigma) = p(-3 < Z < 3) = 0.9973

$$P(|Z| > 3) = 1 - p(|Z| \leq 3) = 0.0027$$

Examples:

If $X$ is normally distributed with mean 3 and standard deviation 2 find.

(i) $P (0 \leq X \leq 4)$

(ii) $P (|X - 3| < 4)$.

Solution:

Given $\mu = 3$, $\sigma = 2$

(i) $P (0 \leq X \leq 4)$

We know that $Z = (X - \mu) / \sigma$

When $X = 0$, $Z = (0 - 3) / 2 = -3 / 2 = -1.5$

When $X = 4$, $Z = (4 - 3) / 2 = 1 / 2 = 0.5$

$P (0 \leq X \leq 4) = P (-1.5 < Z < 0.5)$

$= P (0 < Z < 1.5) + P (0 < Z < 0.5)$

$= 0.4332 + 0.1915$

$= 0.6249$

(ii) $P (|X - 3| < 4) = P (-4 < (X - 3) < 4) \Rightarrow P (-1 < X < 7)$

When $X = -1$, $Z = (-1 - 3)/2 = -4/2 = -2$

When $X = 7$, $Z = (7 - 3)/2 = 4/2 = 2$

$P (-1 < X < 7) = P (-2 < Z < 2)$

$= P (0 < Z < 2) + P (0 < Z < 2)$

$= 2(0.4772) = 0.9544$
6. Beta Distribution

A continuous r.v. X is said to have beta distribution with parameters and if the p.d.f. of x is

\[ f(x; \alpha, \beta) = \frac{1}{\beta(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} \]

For \(0 < x < 1\).

The mean of beta distribution is \(\frac{\alpha}{\alpha + \beta}\) and variance is \(\frac{\alpha \beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)}\). The time X necessary for laying the foundation of a building follows beta distribution.

7. Gamma Distribution

A continuous r.v. X is said to follow a gamma distribution with parameters \(n\) and \(\alpha\) if the p.d.f. of X is

\[ f(x; n, \alpha) = \frac{1}{\alpha \Gamma(n)} x^{n-1} e^{-x/\alpha} \]

For \(x \geq 0\) and where \(n \geq 0, \alpha > 0\).

The mean of the gamma distribution is \(n\) and \(\alpha\), variance is \(n \alpha^2\). It is highly skewed distribution.

Self assessment questions

1. State whether the following statements are true ‘T’ or false ‘F’

   i. The sum of probabilities sometimes will be greater than 1.
   ii. The amount of time you study for an exam is a discrete random variable.
   iv. The Bernoulli distribution has only one parameter ‘p’
   iv. Mean of binomial distribution is ‘npq’.
   v. ‘n’ and ‘p’ are the parameters of binomial distribution.
   vi. If the mean and variance of binomial distribution are 6 and 5, then \(p = 1/6\)
   vii. \(X\) is a poisson variate if \(P<0.1\) and \(n>10\)
   viii. Example of binomial distribution is poisson distribution
   xii. Quartile deviation of normal distribution is \(4/5\sigma\)

   xiii. Mean and standard deviation of a standard normal distribution are ‘1’ or ‘0’

   xiv. Mean, median and mode coincide in normal distribution
CHAPTER 6

Sampling Distributions

We will discuss about the statistical sampling and sampling designs. In different fields of human activity, the decision making process is based on the observations of few units which form a portion of the total population. The process of studying only a portion of the population and making decisions involves risk, the risk of making wrong decisions. This unit deals with the various techniques of drawing samples from the population.

When sampling design is not properly, the estimation or inferences drawn from the sample can go wrong and the managerial decisions taken on the wrong conclusions may lead to loss of time, money and human resource. This may badly affect the reputation of their organization. Hence, the risk involved in using the incorrect sampling design are of primary concerns to investigators.

Universe or Population

Statistical survey or enquiries deal with study various characteristics of unit belonging to a group. The group consisting of all the units is called Universe of Population. The Fig. 7.1 illustrates the population

Types of population

The Fig. 7.2 displays the types of population along with explanation.

<table>
<thead>
<tr>
<th>Finite Population</th>
<th>A population with finite number of units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infinite Population</td>
<td>A population with infinite number of units</td>
</tr>
<tr>
<td>Existent Population</td>
<td>A population of concrete objects i.e. Books in the library</td>
</tr>
<tr>
<td>Hypothetical Population</td>
<td>Throwing a coin infinite number of times</td>
</tr>
</tbody>
</table>
Sample

Sample is a finite subset of a population. A sample is drawn from a population to estimate the characteristics of the population. Sampling is a tool which enables us to draw conclusions about the characteristics of the population. The Fig. 7.3 illustrates the population and sample.

Advantages of sampling

a. The advantages of sampling are:
b. In short time we get maximum information about the population
c. It results in considerable amount of saving of time and labour.
d. The organization and administration of a sample survey in relatively much less.
e. The results obtained are reliable and always possible to attach degree of reliability.
f. There is a possibility of obtaining detailed information. In other words there is a greater scope.
g. In case of infinite population, it is the only available method.
h. If the units are destroyed or affected adversely in the course of investigation, then the only method in sampling.

Sampling theory

The sampling theory is based on the following important laws. The Fig. Below shows the five important laws of sampling theory.

a. Law of statistical regularity
b. Principle of inertia of large numbers
c. Principle of persistence of small numbers
d. Principle of validity
e. Principle of optimization
f.  

1 Law of statistical regularity
The law of statistical regularity states that a group of units chosen at random form a large group tends to possess the characteristics of that large group. Suppose a particular characteristic of the population has a particular shape, then the same characteristics will also follow the same shape in the sample.

2 Principle of inertia of large number

This principle states that ‘other things being equal, as the sample size increases, the results tend to more reliable and accurate. Suppose that the population mean is 25 units. If a sample size of 50 results in average of 24.5 units, then larger sample size of 100 will result in 24.8 units. In other words, larger the sample size, more accurate will be the result.

3 Principle of persistence of small numbers

If some of the units in a population possess markedly distinct characteristics, then it will be reflected in the sample values also. For e.g., if there are 300 blind persons in a population of 10,000 persons, then a sample of hundred will have more or less same proportion of blind persons in it.

Terms used in Sampling Theory

Parameter: Any statistics, like mean, median, calculated from population values are known as parameters of the population and denoted by Greek letters (μ, and σ so on).

Statistics: Any statistics calculated from the sample are known as statistic and are denoted by English letters (x̄, s and so on). Statistic is the parameter of a sample.

Sampling distribution: Sampling distribution consists of all the possible values of a statistic and their respective probabilities for a given sample size.

Principle of validity

A sampling design is said to be valid if it enables us to obtain tests and estimation about population parameters.

Principle of optimization

The principle aims at obtaining a desired level of efficiency at minimum cost or obtaining maximum possible efficiency with given level of cost.

Solved problem: Consider the selection of two numbers form the given five numbers (1, 2, 3, 4, 5). Find the possible combinations and their mean.

Solution: The possible combinations and their average are represented in Table given below:
This gives the mean of sample size 2. We form a distribution of sample mean which can be represented in Table given below:

<table>
<thead>
<tr>
<th>X(Mean)</th>
<th>F(Frequency)</th>
<th>fx</th>
<th>fx²</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5</td>
<td>1</td>
<td>1.5</td>
<td>2.25</td>
</tr>
<tr>
<td>2.0</td>
<td>1</td>
<td>2.0</td>
<td>4.00</td>
</tr>
<tr>
<td>2.5</td>
<td>2</td>
<td>5.0</td>
<td>12.50</td>
</tr>
<tr>
<td>3.0</td>
<td>2</td>
<td>6.0</td>
<td>18.00</td>
</tr>
<tr>
<td>3.5</td>
<td>2</td>
<td>7.0</td>
<td>24.50</td>
</tr>
<tr>
<td>4.0</td>
<td>1</td>
<td>4.0</td>
<td>16.00</td>
</tr>
<tr>
<td>4.5</td>
<td>1</td>
<td>4.5</td>
<td>20.25</td>
</tr>
<tr>
<td>N</td>
<td>10</td>
<td>30</td>
<td>97.50</td>
</tr>
</tbody>
</table>

\[ \text{Mean of distribution} = \frac{\sum fx}{N} \]

\[ \text{Mean of population} = \frac{1 + 2 + 3 + 4 + 5}{5} = 3. \]
The Table represents the sampling distributions of means. We observe that the mean of sample means is equal to population mean.

**Error in Statistics**

The term ‘error’ denotes the difference between population value and its estimate provided by sampling technique. Therefore, the term is not referred in its ordinary sense in statistics.

**Student’s t- distribution**

If random sample $X_1, X_2, ..., X_n$ of size n, with observed values $x_1, x_2, ..., x_n$ is drawn from a normal population having mean $\mu$ and S.D. $\sigma$, the mean $x$ is distributed normally with mean $\mu$ and S.D. $\frac{\sigma}{\sqrt{n}}$ i.e. $\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

Also the variable $Z$, where,

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Is a normal variate with mean 0 and standard deviation 1, i.e. $Z \sim N(0, 1)$.

In practice, the standard deviation $\sigma$ is not known and in such a situation the only alternative left is to, use $s$, the sample estimates of standard deviation $\sigma$. Thus the variate $\sqrt{n} \left(\bar{x} - \mu\right)/s$ is approximately normal provided n is sufficiently large. If n is not sufficiently large, the variate $\sqrt{n} \left(\bar{x} - \mu\right)/s$ is distributed as $t$ and hence

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Where,

$$s^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$$

$T$ is widely used variable and the distribution of $t$ is called student’s t-distribution. This distribution was discovered by W.S. Gosset in 1908. The statistician Gosset is better known by the pseudonym ‘student’ and hence $t$-distribution is called student’s $t$-distribution.

**Properties of ‘t’ distribution**

1. $t$-distribution is unimodal distribution
2. The probability distribution curve is a symmetrical about the line $t=0$.
3. It is bell shaped curve just like a normal curve with its tails a little higher above the abscissa than the normal curve. Its spread increases as degrees of freedom ‘K’ decreases.
This means that for the same value of t-variate and x, the normal variate, the area beyond t is larger than the area beyond x.

4. t-distribution has only one parameter k, the degrees of freedom equal to (n-1).

5. The constants of t-distribution are as follows:
   (Mean) $\mu = 0$ for $k \geq 2$.
   (Variance) $\sigma^2 = \frac{k}{k - 2}$ for $k \leq 3$
   (Skewness) $\alpha_3 = 0$ for $k > 4$.
   (Kurtosis) $\alpha_4 = \frac{3(k - 2)}{(k - 4)}$ for $k \leq 5$.

6. The area under t-distribution curve for $t < t'$ is determined by the equation.

$$F_t(t) = p(t - t') = \int_{-\infty}^{t'} f(t) dt \quad 8.4$$

Students and other readers need not integrate actually for the area as the table of area under the curve for different values of ‘t’ are available and vice-versa. Equation (8.4) is given to acquaint the reader with the know-how of getting the area if t is given and the value of t if area is given.

7. t-distribution tends to normal distributions as k increases. For practical purposes, t is taken an equivalent to the normal distribution provided $k \geq 30$.

8. Moment generating function for t-distribution does not exist.

The t distribution has tremendous utility in testing of hypothesis about one population mean or about equality of two population means when standard deviation is not known.

**Chi-Square distribution**

So far, we have been discussing the mean obtained from all possible samples, or large number of samples drawn from a normal population, distributed with mean $\mu$ and variance $\sigma^2/n$. Now we are interested in knowing the distribution of sample variances $s^2$ of these samples. Consider a random sample $X_1, X_2, \ldots, X_n$ of size n. Let the observations of this sample be denoted by $x_1, x_2, \ldots, x_n$. We know that the variance.

$$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

for $i = 1, 2, \ldots n$.

or

$$\sum (x_i - \bar{x})^2 = (n-1)s^2 = ks^2$$

where, $k = (n-1)$
A quantity $k^2/w$, which is a pure number, is defined as $\chi^2_k$. Now we will give the distribution of the random variable $\chi^2_k$, which was first discovered by Helmert in 1876 and later independently given by Karl Pearson in 1900. Another way to understand chi-square is: if $X_1, X_2, \ldots, X_n$ are $n$ independent normal variates with mean zero and variance unity, the sum of squares of these variates is attributed as chi-square with n df. The chi-square distribution was discovered mainly as chi-square with n d.f. The chi-square distribution was discovered mainly as a measure of goodness of fit in case of frequency distribution, i.e. whether the observed frequencies follow a postulated distribution or not. The probability density function (P.d.f.) of $\chi^2$–variate is,

$$f_X(\chi^2) = \frac{1}{2^{k/2} \Gamma(k/2)} (\chi^2)^{k/2-1} e^{-\frac{1}{2}\chi^2}$$

**Properties of Chi-square Distribution**

i. The whole chi-square distribution curve lies in the first quadrant since the range of $\chi^2$ is from 0 to $\infty$.

ii. From the density function, it is evident that $\chi^2$-distribution has only one parameter $k$, the degrees of freedom for $\chi^2$.

Thus, the shape of the probability density curve mainly depends on the parameter $k$. The shape of the curves for four different degrees of freedom say, $k=2, k=7, k=12, k=20$ are given below:

iii. Chi square distribution curve is highly positive skewed.

iv. It is unimodal curve and its mode is at the point $\chi^2 = (k-1)$.

v. The shape of the curve varies immensely especially when $k$ is small. $k=1$ and 2, it is just like a hyperbola.

vi. Chi square distribution is completely defined by one parameter ‘$k$’, which is known as degree of freedom of chi-square distribution.

vii. The constants for chi-square distributions are as follows

(mean) $\mu = k$

(variance) $\sigma^2 = 2k$

Skewness $\alpha_1 = 2 \left( \frac{2}{k} \right)^{1/3}$

viii. The moment generating function for chi-square distribution is,

$\phi_\chi(t) = (1 - 2t)^{-k/2}$

ix. $r$-the raw moment of chi square distribution is,

$\mu'_r(t) = \frac{2^r \Gamma \left( \frac{k}{2} + r \right)}{\Gamma \left( k/2 \right)}$
Putting $r = 1$,

$$\mu' = \frac{2 \Gamma \left( \frac{k}{2} + 1 \right)}{\Gamma k / 2} = k$$

$$= 2, \quad \mu'_2 = \frac{2^2 \Gamma \left( \frac{k}{2} + 2 \right)}{\Gamma k / 2} = 4 \left( \frac{k}{2} + 1 \right) = k (k + 2)$$

Again

$$\mu^2 = \mu'_2 - (\mu'_1)^2 = k (k + 2) - k^2 = 2k$$

and so on.

x. It can be shown that for large degrees of freedom say; $k \geq 100$, the variable $\left( \sqrt{2} \chi^2 - \sqrt{2k - 1} \right)$ is distributed normally with mean 0 and variance 1.

xi. If $\chi^2_1$ and $\chi^2_2$ are two independent chi-squares with degrees of freedom $k_1$ and $k_2$, respectively, their sum $(\chi^2_1 + \chi^2_2)$ will be distributed as chi-square with $(k_1 + k_2)$ d.f. This additive property of independent chi-squares, hold good for any number of chi-squares i.e. if there are $m$ independent chi-squares with $k_1, k_2, ..., k_m$ d.f, respectively, the sum $\sum_i \chi^2_i$ (i=1, 2, . . . m) will be distributed as chi-square with $\sum_i k_i$ d.f.

**F distribution**

It is evidence that the ratio of two independent sample variances is denoted by $F$. Now we will consider the distribution of the ratio of two-sample variance in another manner and based on this new approach, the distribution of F is given. The distribution of F was worked out by G.W. Snedecor.

Consider two normal populations $N (\mu_1, \sigma^2_1)$ and $N (\mu_2, \sigma^2_2)$. An independent sample of size $n_1$ is drawn from a population $N (\mu_1, \sigma^2_1)$ and of size $n_2$ from a population $N (\mu_2, \sigma^2_2)$. Let the sample variance be $s^2_1$ and $s^2_2$, respectively. From the chi-square distribution theory, we know that $k_1 s^2_1 / \sigma^2$ is distributed as chi-square with $k_1$ d.f. i.e.

$$\frac{k_1 s^2_1}{\sigma^2_1} \sim \chi^2_1$$

Where, $\chi^2_1$ has d.f. $k_1 = (n_1 - 1)$
\[ \frac{k_2 s^2}{\sigma^2_2} \sim \chi^2_2 \]

Where, \( \chi^2_2 \) has d.f. \( k = (n_2-1) \)

From the above two equations, we can write

\[
\frac{s^2_1/\sigma^2_1}{s^2_2/\sigma^2_2} = \frac{\chi^2_1/k_1}{\chi^2_2/k_2} = F_{k_1,k_2}
\]

Here, \( k_1 \) and \( k_2 \) are called the degrees of freedom of \( F \). Dividing \( s^2_u \), \( (u=1,2) \) by its corresponding population variance standardizes the sample variance, in the sense that on the average both the numerator and denominator approach 1. Now we may be interested in testing the hypothesis that both the normal population have the same variance, i.e. \( \sigma^2_1 = \sigma^2_2 \).

Under this hypothesis.

\[
\frac{s^2_1}{s^2_2} = F_{k_1,k_2}
\]

As a norm, the greater sample variance is taken as the numerator.

From above equation, it is apparent that the ratio of two independent chi-squares is distributed as \( F \) and reveals that under the hypothesis \( \sigma^2_1 = \sigma^2_2 \), the ratio of two independent sample variances is distributed as \( F \). The probability density function of \( F \)-distribution is,

\[
f_{k_1,k_2}(F) = \frac{(k_1/k_2)^{k_1/2}}{B(k_1/2, k_2/2)} \left( 1 + \frac{k_1}{k_2} F \right)^{-(k_1+k_2)/2} \]

\[ 0 \leq F < \infty \]

Obviously, \( F \) is always a positive number and \( F \)-distribution curve wholly lies in the first quadrant. There are two parameters of \( F \)-distribution namely, \( k_1 \) and \( k_2 \). Hence, the shape of \( F \)-distribution curve depends on \( k_1 \) and \( k_2 \).

**Properties of \( F \) – Distribution**

1. \( F \)-distribution curve extends on abscissa from 0 to \( \infty \).
2. It is unimodal curve and its mode lies on the point
\[ F = \frac{k_2(k_1 - 2)}{k_2(k_1 + 2)} \]

Which is always less than unity.

3. \( F \)-distribution curve is positive skew curve. Generally, the \( F \)-distribution curve is highly positive skewed when \( k_2 \) is small.

4. The constants of \( F \)-distribution are,

   \begin{align*}
   \text{(mean)} & \quad \mu = \frac{k_2}{k_2 - 2} \quad \text{for} \quad k_2 \geq 3 \\
   \text{(variance)} & \quad \sigma^2 = \frac{2k_1^2(k_1 + k_2 - 2)}{k_1(k_2 - 2)(k_2 - 4)} \quad \text{for} \quad k_2 \geq 5.
   \end{align*}

5. There exists a very useful relation for interchange of degree of freedom \( k_1 \) and \( k_2 \) i.e.

   \[ F_{1-\alpha(k_1,k_2)} = \frac{1}{F_\alpha(k_2,k_1)} \]

6. The moment generating function of \( F \)-distribution does not exist.

F-distribution is a very popular and useful distribution because of its utility in testing of hypothesis about the equality of several population means, two population variances and several regression coefficients in multiple regressions etc. As a matter of fact, F-test is the backbone of analysis of variance.
CHAPTER 7

Test of Significance

This involves certain amount of risk. This amount of risk is termed as level of significance. When the hypothesis is accepted, we consider it a non-significant result and if the reverse situation occurs, it is called a significant result. A test is defined as “A statistical test is a procedure governed by certain rules, which leads to take a decision about the hypothesis for its acceptance or rejection on the basis of sample values.

Types of hypothesis

A hypothesis is an assertion or conjecture about the parameter(s) of population distribution(s).

Null Hypothesis:

A hypothesis which is to be actually tested for acceptances or rejection is termed as null hypothesis. It is denoted by \( H_0 \)

Alternative Hypothesis:

It is a statement about the population parameter or parameters, which gives an alternative to the null hypothesis \( (H_0) \), within the range of pertinent values of parameter, i.e., If \( H_0 \) is accepted, what hypothesis is to be rejected and vice versa. An alternative hypothesis is denoted by \( H_1 \) or \( H_A \). The idea of alternative hypothesis was originated by Nayman.

Simple and Composite Hypothesis:

If the statistical hypothesis completely specifies the distribution, it is called simple hypothesis, otherwise it is called a composite hypothesis.

Two types of Errors

After applying a test, a decision is taken about the acceptance or rejection of null hypothesis vis-a-vis the alternative hypothesis. There is always some possibility of committing an error in taking a decision about the hypotheses. These errors can be of two types.

Type I error: Reject null hypothesis \( (H_0) \) when it is true.

Type II error: Accept null hypothesis \( (H_0) \) when it is false.
Level of significance:

It is the quantity of risk of the type I error which we are ready to tolerate in making a decision about $H_0$. In other words, it is the probability of type I error which is tolerable.

P-Value concept:

Another approach is to find out the P-value at which $H_0$ is significant, i.e., to find out the smallest level $\alpha$ of at which $H_0$ is rejected. In this situation, it is not inferred whether $H_0$ is accepted or rejected at level 0.05 or 0.01 or any other level. But the statistician only gives the smallest level $\alpha$ at which $H_0$ is rejected.

Critical Region (C.R.)

A statistic is used to test the hypothesis $H_0$. The test statistic follows some known distribution. In a test, the area under the probability density curve is divided into two regions viz., the region of acceptance and the region of rejection. The region of rejection is the region in which $H_0$ is rejected. It means that if the value of test statistics lies in this region, $H_0$ will be rejected. The region of rejection is called critical region.

Randomized Test:

A randomized test (T) is one in which no test statistic is used. The decision about the rejection of $H_0$ is taken, if satisfies some predecided criterion.

Non-randomized test:

A test T of a hypothesis H is said to be non-randomized if the hypothesis H is rejected on the basis that a test statistic belongs to the critical region $C_r$, i.e. $\Psi (X_1, X_2, \ldots, X_n) \in C_r$.

Degrees of Freedom (d.f.):

It is apparent from the discussion made so far that in a test of hypothesis, a sample is drawn from the population of which the parameter is under test. The size of sample varies since the depends either on the experimenter or on the resources available.

Definition: Degree of freedom is the number of independent observations in a set.
Student’s t-Test

We shall make use of t-distribution in testing of hypothesis about the population mean or means.

Suppose a small random sample \((X_1, X_2, \ldots, X_n)\) of size \(n\) has been drawn from a normal population having mean \(\mu\) and variance \(\sigma^2\) which are unknown. We want to test the hypothesis.

\[
H_0: \mu = \mu_0 \quad \text{against} \quad H_1: \mu \neq \mu_0
\]

Where, \(\mu_0\) is some assumed value considered fit for \(\mu\). Let the observed values on random sample \((X_1, X_2, \ldots, X_n)\) be \((x_1, x_2, \ldots, x_n)\). Statistic \(t\) is given as,

\[
t = \frac{\bar{x} - \mu_0}{s}
\]

For the sample values, expression for \(t\) is,

\[
t = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s}
\]

Where, the sample mean, \(s\) is the standard deviation of the sample, \((n-1)\) in suffix indicates the d.f. of \(t\).

On substituting formula for \(s\), it can be written as-

\[
t = \left(\bar{x} - \mu_0\right) \sqrt{\frac{n(n-1)}{\sum (x_i - \bar{x})^2}}
\]

Where, \(i = 1, 2, \ldots, n\).

Assumptions about t-Test: t-test based on the following five assumptions

1. The random variable \(X\) follows normal distribution. In other word the random sample has been drawn from a normal population.
2. All observations in the sample are independent.
3. The sample size is not large. There is no hard-and-fast rule which can be given to call a sample large. But, as a practice, a sample size of 30 or more is considered a large sample. At the same time one should note that at least five observations are desirable for applying a t-test.
4. The assumption value \(\mu_0\) of the population mean is the correct value.
5. The sample values are correctly taken and recorded.
Example:

A breeder claims that his variety of cotton contains, at the most 40 per cent Lint in seed cotton. Eighteen samples of 100 grams each were taken, and after ginning the following quantity of lint was found in each sample.

<table>
<thead>
<tr>
<th>Sample No.</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>Quantity of lint in 100 g sample</td>
<td>36.3</td>
<td>37.0</td>
<td>36.6</td>
<td>37.5</td>
<td>37.5</td>
<td>37.9</td>
<td>37.8</td>
<td>36.9</td>
<td>36.7</td>
</tr>
<tr>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>38.5</td>
<td>37.9</td>
<td>38.8</td>
<td>37.5</td>
<td>37.1</td>
<td>37.0</td>
<td>36.3</td>
<td>36.7</td>
<td>35.7</td>
<td></td>
</tr>
</tbody>
</table>

To check the breeder’s claim, a t-test is performed as under. Here we have to test:

\[ H_0 : \mu = 40 \] against \[ H_1 : \mu < 40 \]

To test \( H_0 \), the test statistic is

\[ t_{n-1} = \frac{\sqrt{n} (\bar{x} - \mu_0)}{s} \]

Given \( \mu_0 = 40 \)

Now we compute \( \bar{x} \) and \( s \).

\[ \bar{x} = \frac{669.7}{18} = 37.206 \]

\[ s^2 = \frac{1}{n-1} \left\{ \sum x_i^2 - \left( \sum x_i \right)^2 / n \right\} \]

\[ s^2 = \frac{1}{17} \left\{ 24927.33 - \left( \frac{669.7}{} \right)^2 / 18 \right\} \]

\[ = \frac{10.77}{17} = 0.633 \]

\[ s = 0.796 \]

\[ t = \frac{\sqrt{n} (\bar{x} - \mu_0)}{s} \]
The table value of $t$ at the prefixed $\alpha = 0.01$ and 17 d.f. is 2.567. Here $t < -2.567$. Hence, $H_0$ is rejected. It means that the average % of lint in this cotton variety is less than 40 per cent.

**Chi-Square test**

Chi square test is one of the most commonly used tests of significance. The chi-square test is applicable to test the hypothesis of the variance of a normal population, goodness of fit of the theoretical distribution to observed frequency distribution, in a one way classified having k-categories. It is also applied for the test of independence of attributes, when the frequencies are presented in two-way classification called the contingency table. The chi-square test dates back to 1900, when Karl Pearson used it for frequency used test in genetics into K mutually exclusive categories. It is also a frequently used test in genetics, were one test whether the observed frequencies in different crosses agree with the expected frequencies or not. Now we given chi-square test of various hypothesis in sufficient details one by one.

**Test of hypothesis for population variance**

Suppose, on the basis of previous knowledge, we have preconceived value, $\sigma_0^2$. Of variance of a normal population. Draw a random sample of size $n$ ($n < 30$) from this population. On the basis of $n$ sample observations ($x_1, x_2, ..., x_n$) the postulated value $\sigma_0^2$ of the population variance $\sigma^2$ is to either be substantiated or refuted with the help of statistical test. For this the hypothesis

$$H_0 : \sigma^2 = \sigma_0^2 \text{ vs } H_1 : \sigma^2 \neq \sigma_0^2$$

Is tested by the statistic

$$\chi^2 = \frac{\sum (x_i - \bar{x})^2}{\sigma_0^2}$$

$$i = 1, 2, \ldots, n$$
Test of Goodness of Fit

To test the assertion of how closely the actual distribution approximates to a particular theoretical distribution, chi-square test is appropriate to assume that the population is distributed normally is a common practice and hence we explain the test of goodness of fit of normal population first.

Let there be \( k \) class intervals and the corresponding frequencies be \( f_1, f_2,\ldots,f_i \). The area of normal curve within each interval is found from the table of area under the normal curve. On multiplying this area by the total of frequencies, we get the expected (theoretical) frequencies. In this way, the expected frequency for each interval is obtained. Since the tables are provided for standard normal curve, we first change each limit of class interval into a standard normal deviate by using the formula:

\[
Z = \frac{x - \bar{x}}{s}
\]

Where, \( \bar{x} \) is the sample mean and \( s \) is the sample standard deviation.

**Example**: The data regarding supplemental security income (SSI) programme to escape from poverty, the poor people over 65 years are enrolled up to 1975 in an area are as follows:

<table>
<thead>
<tr>
<th>Annual Benefit ($)</th>
<th>Mid values (x)</th>
<th>No of individuals in own home (f)</th>
<th>Expected frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-100</td>
<td>50</td>
<td>7</td>
<td>7.62</td>
</tr>
<tr>
<td>100-250</td>
<td>175</td>
<td>9</td>
<td>6.32</td>
</tr>
<tr>
<td>250-500</td>
<td>375</td>
<td>19</td>
<td>15.26</td>
</tr>
<tr>
<td>500-750</td>
<td>625</td>
<td>12</td>
<td>16.17</td>
</tr>
<tr>
<td>750-1000</td>
<td>875</td>
<td>8</td>
<td>11.52</td>
</tr>
<tr>
<td>1000-1250</td>
<td>1125</td>
<td>5</td>
<td>5.24</td>
</tr>
<tr>
<td>1250-1500</td>
<td>1375</td>
<td>4</td>
<td>1.88</td>
</tr>
</tbody>
</table>

Whether the given frequency distribution follows the normal distribution or not can be tested by the chi square test.

The expected frequencies are worked out in the following manner.

To obtain the standard normal deviates, we calculate the mean \( \bar{x} \) and standard deviation \( s \).
The calculations are

\[ \sum f_i = 64, \sum f_i x_i = 34675, \text{ and } \sum f_i x_i^2 = 27668125 \]

\[ \bar{x} = \frac{34675}{64} = 541.8 \]

\[ s^2 = \frac{1}{63} \left\{ 27668125 - (541.8)^2 \times 64 \right\} \]

\[ = \frac{1}{63} \times 8881102 = 140969.87 \]

\[ s = 375.46 \]

The values of standard normal deviates \((X - \bar{x})/s\) are,

\[ Z_1 = \frac{0 - 541.8}{375.46} = -1.44 \]

\[ Z_2 = \frac{100 - 541.8}{375.46} = -1.18 \]

Similarly,

\[ Z_3 = 0.78, Z_4 = -0.11, Z_5 = 0.55, Z_6 = 1.22, Z_7 = 1.80, Z_8 = 2.55 \]

From Table IV, the area between \(Z_1\) and \(Z_2\)

\[ = 0.42507 - 0.38100 = 0.04407 \]

To calculate the expected frequency in the lowest class 0-100, the area of the left of 0 (-\(\infty\) to 0) i.e. an area = 0.5 - 0.42507 = 0.07493 will also be included.

Hence, the area, to calculate the expected frequency for the lowest class = 0.07493 + 0.04407 = 0.11900

Area between \(Z_2\) and \(Z_3\) = 0.38100 - 0.28230 = 0.09870
Area between \(Z_3\) and \(Z_4\) = 0.28230 - 0.04383 = 0.23850
Area between \(Z_4\) and \(Z_5\) = 0.04383 + 0.20884 = 0.25267
Area between \(Z_5\) and \(Z_6\) = 0.38877 - 0.20884 = 0.17993
Area between \(Z_6\) and \(Z_7\) = 0.47062 - 0.38877 = 0.08185
Area between \(Z_7\) and \(Z_8\) = 0.49461 - 0.47062 = 0.02399

To calculate the expected frequency in the upper most class (1250-1500) the area to the right of 1500 (1500 to \(\infty\)) i.e. an area
\[ 0.5 - 0.49461 = 0.00539 \]

Will also be included.

Hence the area, to calculate the expected frequency for the upper most class = 0.02399 + 0.00539 = 0.02938

The expected frequencies are calculated as,

\[
\begin{align*}
  f_1 &= 0.11900 \times 64 = 7.62, \\
  f_2 &= 0.09870 \times 64 = 6.32 \\
  f_3 &= 0.23850 \times 64 = 15.26, \\
  f_4 &= 0.25267 \times 64 = 16.17 \\
  f_5 &= 0.17993 \times 64 = 11.52, \\
  f_6 &= 0.08185 \times 64 = 5.24 \\
  f_7 &= 0.02938 \times 64 = 1.88
\end{align*}
\]

These frequencies are depicted in the last column of the frequency distribution.

Since the expected frequency of the upper most class is <5, this class is collapse with the preceding class. In this way, the number of classes is reduced to 6. This makes the chi-square test statistics good indicator of the validity of the hypothesis under test.

\[
\chi^2 = \frac{(7 - 7.62)^2}{7.62} + \frac{(9 - 6.32)^2}{6.32} + \frac{(19 - 15.26)^2}{15.26} + \frac{(12 - 16.17)^2}{16.17} + \frac{(8 - 11.52)^2}{11.52} + \frac{(9 - 7.12)^2}{7.12}
\]

\[= 0.0504 + 1.1364 + 0.9166 + 1.0753 + 1.0756 + 0.4964 \]

\[= 4.751 \]

**Computation of the chi-squared statistic**

The first step in computing the Chi-squared statistic is the computation of the contingency table. The preceding table is reproduced below.

<table>
<thead>
<tr>
<th>AIDS</th>
<th>SEXREF</th>
<th>Males</th>
<th>Females</th>
<th>Both</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AIDS</strong></td>
<td></td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td><strong>NO AIDS</strong></td>
<td></td>
<td>3</td>
<td>16</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>18</td>
<td>5</td>
</tr>
</tbody>
</table>

**Teaching Manual on “Statistical Methods for Applied Sciences” - Ravi R Saxena, M.L. Lakhera and Roshan Bhardwaj, Department of Agricultural Statistics and SS (L), IGKV, Raipur – 492 012 (Chhattisgarh)**
The 2\textsuperscript{nd} step in computing the Chi-squared statistic is the computation of the expected cell frequency for each cell. This is accomplished by multiplying the marginal frequencies for the row and column (row and column totals) of the desired cell and then dividing by the total number of observations. The formula for computation can be represented as

\[
\text{Expected Cell Frequency} = \frac{(\text{Row Total} \times \text{Column Total})}{N}
\]

For example, computation of the expected cell frequency for Males with AIDS would proceed as follows

\[
\text{Expected Cell Frequency} = \frac{(\text{Row Total} \times \text{Column Total})}{N} = \frac{(9 \times 7)}{30} = 2.1
\]

Using the same procedure to compute all the expected cell frequencies results in the table given below

The 3\textsuperscript{rd} step is to subtract the expected cell frequency from the observed cell frequency for each cell. This value gives the amount of deviation or error for each cell. Adding these to the preceding table results in the following:
### Note:

1. The sum of the expected row total is the same as the sum of the observed row totals; the same holds true for the column totals.
2. The sum of the Observed - Expected for both the rows and columns equals zero.

**The 4th step** the difference computed in the last step is squared.

### Table

<table>
<thead>
<tr>
<th>SEXPREF</th>
<th>Males</th>
<th>Females</th>
<th>Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>4</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Expected</td>
<td>2.1</td>
<td>5.4</td>
<td>1.5</td>
</tr>
<tr>
<td>$O-E$</td>
<td>1.9</td>
<td>-3.4</td>
<td>1.2</td>
</tr>
<tr>
<td>$(O-E)^2$</td>
<td>3.61</td>
<td>11.56</td>
<td>2.25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SEXPREF</th>
<th>Males</th>
<th>Females</th>
<th>Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>3</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>Expected</td>
<td>4.9</td>
<td>12.6</td>
<td>3.5</td>
</tr>
<tr>
<td>$O-E$</td>
<td>-1.9</td>
<td>3.4</td>
<td>1.5</td>
</tr>
<tr>
<td>$(O-E)^2$</td>
<td>3.61</td>
<td>11.56</td>
<td>2.25</td>
</tr>
</tbody>
</table>

The 5th step: each of the squared differences is then divided by the expected cell frequency for each cell, resulting in the following table.

### Table

<table>
<thead>
<tr>
<th>SEXPREF</th>
<th>Males</th>
<th>Females</th>
<th>Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>4</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Expected</td>
<td>2.1</td>
<td>5.4</td>
<td>1.5</td>
</tr>
<tr>
<td>$O-E$</td>
<td>1.9</td>
<td>-3.4</td>
<td>1.2</td>
</tr>
<tr>
<td>$(O-E)^2$</td>
<td>3.61</td>
<td>11.56</td>
<td>2.25</td>
</tr>
<tr>
<td>$(O-E)^2/E$</td>
<td>1.72</td>
<td>2.14</td>
<td>1.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SEXPREF</th>
<th>Males</th>
<th>Females</th>
<th>Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>3</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>Expected</td>
<td>4.9</td>
<td>12.6</td>
<td>3.5</td>
</tr>
<tr>
<td>$O-E$</td>
<td>-1.9</td>
<td>3.4</td>
<td>1.5</td>
</tr>
<tr>
<td>$(O-E)^2$</td>
<td>3.61</td>
<td>11.56</td>
<td>2.25</td>
</tr>
<tr>
<td>$(O-E)^2/E$</td>
<td>0.74</td>
<td>0.92</td>
<td>0.64</td>
</tr>
</tbody>
</table>
The chi-square statistic is computed by summing the last row of each cell in the preceding table, the formula being represented by:

\[ \chi^2_{\text{obs}} = \sum_{\text{cells}} \frac{(O - E)^2}{E} \]

This computation for the example table would result in the following:

\[ \chi^2_{\text{obs}} = 1.72 + 2.14 + 1.50 + .74 + .92 + .64 = 7.66 \]

**Note:**

- This value is within rounding error of the value for Chi-square computed by the computer in an earlier section of this chapter.

**Interpretation:**

The cell frequencies may be guided by the amount each cell contributes to the chi-squared statistic, as seen in the \((O - E)^2/E\) value. In general, the larger the difference between the observed and expected values, the greater this value. In the example data, it can be seen that the homosexual males had a greater incidence of Aids (Observed = 4, Expected = 2.1) than would be expected by chance alone, while heterosexual had a lesser incidence (Observed = 2, Expected = 5.4). This sort of evidence could direct the search for the causes of Aids.

<table>
<thead>
<tr>
<th>SEXPREF</th>
<th>Males</th>
<th>Females</th>
<th>Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>4</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Expected</td>
<td>(O - E)^2/E</td>
<td>2.1</td>
<td>5.4</td>
</tr>
<tr>
<td></td>
<td>1.72</td>
<td>2.14</td>
<td>1.5</td>
</tr>
<tr>
<td>Observed</td>
<td>3</td>
<td>1.6</td>
<td>2</td>
</tr>
<tr>
<td>Expected</td>
<td>(O - E)^2/E</td>
<td>4.9</td>
<td>12.6</td>
</tr>
<tr>
<td></td>
<td>7.4</td>
<td>9.2</td>
<td>6.4</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>18</td>
<td>5</td>
</tr>
</tbody>
</table>

If the value of the observed chi-square statistic is less than the expected value, then the model of no effects cannot be rejected and the table is not significant. It can be
said that no effects were discovered. In this case an interpretation of the cell frequencies is not required, because the values could have been obtained by chance alone.

**Yates' correction for continuity;**

In statistics, Yates' correction for continuity (or Yates’ chi-square test) is used in certain situations when testing for independence in a contingency table.

Correction for approximation error

Using the chi-squared distribution to interpret Pearson's chi-squared statistic requires one to assume that the discrete probability of observed binomial frequencies in the table can be approximated by the continuous chi-squared distribution. This assumption is not quite correct, and introduces some error.

To reduce the error in approximation, Frank Yates, an English statistician, suggested a correction for continuity that adjusts the formula for Pearson's chi-square test by subtracting 0.5 from the difference between each observed value and its expected value in a $2 \times 2$ contingency table. This reduces the chi-square value obtained and thus increases its $p$-value.

The effect of Yates' correction is to prevent overestimation of statistical significance for small data. This formula is chiefly used when at least one cell of the table has an expected count smaller than 5. Unfortunately, Yates' correction may tend to overcorrect. This can result in an overly conservative result that fails to reject the null hypothesis when it should. So it is suggested that Yates' correction is unnecessary even with quite low sample sizes.

The following is Yates' corrected version of Pearson's chi-squared statistic:

$$\chi^2 = \sum \frac{(O_i - E_i) - 0.5)^2}{E_i}$$

Where:

$O_i$ = an observed frequency

$E_i$ = an expected (theoretical) frequency, asserted by the null hypothesis

$N$ = number of distinct events

**Application of chi-square test**

$\chi^2$ have several applications. They are as follows,-
1. It has goodness of fit.
2. To test if the hypothetical value of population variance \( \sigma^2 = \sigma^2_0 \) (say)
3. To test the independence of attributes.
4. To test the homogeneity of population correlation coefficient.
5. To test the homogeneity of the independent estimation of the population correlation coefficient.

Example:

A sample analysis of examination of result of 200 MBA's was made was found that 46 students had failed, 68 secured 3rd division, 62 secure 2nd division and rest were blessed as 1st division. Is this figure match the figure of 4:3:2:1? In other words the given data are matched with general exam result which is in the ratio of 4:3:2:1 for various categories respectively.

Solution. Given different marks along with the division of 200 MBA students

To find whether the result match with the given data

Setting hypothesis

Null hypothesis \((H_0)\) = the observed figures match with hypothetical value which is in the ratio of 4:3:2:1 for various categories.

Alternate hypothesis \((H_1)\) = the observed figure did not match with hypothetical value.

Level of significance = 0.05 or 5% level of significance.

Formula used.

\[ \chi^2 = \sum_{i=1}^{n} \frac{(O - E)^2}{E} \] with 1 d.f. (d.f. = no. of parameters - no. of restriction imposed on it.)
Observation table

<table>
<thead>
<tr>
<th>Category</th>
<th>Frequency</th>
<th>Expected (Ei)</th>
<th>( \frac{(O_i - E_i)^2}{E_i} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failed</td>
<td>46</td>
<td>80</td>
<td>14.45</td>
</tr>
<tr>
<td>3rd div</td>
<td>68</td>
<td>60</td>
<td>1.067</td>
</tr>
<tr>
<td>2nd div</td>
<td>62</td>
<td>40</td>
<td>12.1</td>
</tr>
<tr>
<td>1st div</td>
<td>24</td>
<td>20</td>
<td>0.8</td>
</tr>
<tr>
<td>Total</td>
<td>28.417</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Answer- calculated value of \( \chi^2 \) is 28.417

Conclusion: - Since the calculated value of \( \chi^2 \) i.e. 28.417 is greater than the table value i.e. 7.815 at n-1=3 d.f. therefore we reject the null hypothesis that there is no difference between the declared result and the expected ratio 4:3:2:1. Hence we may conclude that the data are not matched the general examination results.

Example

The demand for the particular spare part in a factory was found to vary from day to day. In a sample study the following information was obtained:

<table>
<thead>
<tr>
<th>Days</th>
<th>Mon.</th>
<th>Tue.</th>
<th>Wed.</th>
<th>Thurs</th>
<th>Fri.</th>
<th>Sat.</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of parts</td>
<td>1124</td>
<td>1125</td>
<td>1110</td>
<td>1120</td>
<td>1126</td>
<td>1115</td>
</tr>
</tbody>
</table>

Demanded

Solution. Given: The no of spare parts in each day of the week. And the values of \( \chi^2 \) significance at 5, 6, 7, d.f. are respectively 11.07, 12.59, and 14.07.

Setting hypothesis.

Null hypothesis. \( (H_0) \): the no. of parts demanded does not depend upon the days of the week.
Alternate hypothesis \( (H_1) \) = the no. of parts demanded depend upon the days of the week.

Level of significance = 0.05 or 5\% level of significance.

Under the null hypothesis, the expected frequencies of the spare part demanded on each day would be:

\[
\frac{1}{6} (1124 + 1125 + 1110 + 1120 + 1126 + 1115) = \frac{6720}{6} = 1120
\]

Formula used- \( \chi^2 = \frac{(O_i - E_i)^2}{E_i} \)

Observation table for \( \chi^2 \)

<table>
<thead>
<tr>
<th>Category</th>
<th>Frequency</th>
<th>( \frac{(O_i - E_i)^2}{E_i} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed (O(_i))</td>
<td>Expected (E(_i))</td>
<td></td>
</tr>
<tr>
<td>Mon.</td>
<td>1124</td>
<td>1120</td>
</tr>
<tr>
<td>Tue.</td>
<td>1125</td>
<td>1120</td>
</tr>
<tr>
<td>Wed.</td>
<td>1110</td>
<td>1120</td>
</tr>
<tr>
<td>Thurs.</td>
<td>1120</td>
<td>1120</td>
</tr>
<tr>
<td>Fri.</td>
<td>1126</td>
<td>1120</td>
</tr>
<tr>
<td>Sat.</td>
<td>1115</td>
<td>1120</td>
</tr>
<tr>
<td>Total</td>
<td>6720</td>
<td>6720</td>
</tr>
</tbody>
</table>

The calculated value of \( \chi^2_{0.05} \) for 5 d.f. = 0.179.

Conclusion:- since the calculated value for chi-square is much more less than table value thus it is not significance at 5\% level of significance and n-1=5 d.f. hence we may accept the null hypothesis and conclude that the number of parts demanded are same over the six-day period.
F-test

A large number of surveys or experiments are conducted to draw conclusions about the effect of certain factors or treatments. Observations are taken pertaining to the character under study. F-test is used either for testing the hypothesis about the equality of two population variances or the equality of two or more population means. The equality of two population means has been dealt with t-test. Besides a t test, we can also apply an F-test for testing equality of two population means.

Example: Life expectancy in 9 regions of India in 2000 and in 11 regions of India in 2010 was as given in the table below:

<table>
<thead>
<tr>
<th>Regions</th>
<th>Life expectancy (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2000</td>
</tr>
<tr>
<td>1</td>
<td>42.7</td>
</tr>
<tr>
<td>2</td>
<td>43.7</td>
</tr>
<tr>
<td>3</td>
<td>34.0</td>
</tr>
<tr>
<td>4</td>
<td>39.2</td>
</tr>
<tr>
<td>5</td>
<td>46.1</td>
</tr>
<tr>
<td>6</td>
<td>48.7</td>
</tr>
<tr>
<td>7</td>
<td>49.4</td>
</tr>
<tr>
<td>8</td>
<td>45.9</td>
</tr>
<tr>
<td>9</td>
<td>55.3</td>
</tr>
<tr>
<td>10</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
</tr>
</tbody>
</table>

It is desired to confirm, whether the variation in life expectancy in various regions in 2000 and 2010 is same or not.

Let the population in 1900 and 1970 be considered as \( N(\mu_1, \sigma_1^2) \) and \( N(\mu_2, \sigma_2^2) \), respectively.
The hypothesis

\[ H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{against} \quad H_1 : \sigma_1^2 \neq \sigma_2^2 \]

can be tested by F test.

First we calculate \( s_1^2 \) and \( s_2^2 \).

\[ \sum_{i} x_{1i} = 405, \ x_{1i}^2 = 18527.78 \]

\[ s_1^2 = \frac{1}{8} \left( 18527.78 - \frac{(405)^2}{9} \right) \]

\[ = \frac{302.78}{8} = 37.848 \]

\[ s_2^2 = \frac{1}{10} \left( \sum_{j} x_{2j}^2 - \frac{(\sum_{j} x_{2j})^2}{11} \right) \]

\[ \sum_{j} x_{2j} = 598.5, \ \sum_{j} x_{2j}^2 = 32799.91 \]

\[ s_2^2 = \frac{1}{10} \left( 32799.91 - \frac{(598.5)^2}{11} \right) \]

\[ = \frac{236.07}{10} = 23.607 \]

The test statistic,

\[ F = \frac{s_1^2}{s_2^2} \]

\[ = \frac{37.848}{23.607} = 1.603 \]

**Self assessment questions**

State whether the following statements are true ‘T’ or false ‘F’

i. Population is the aggregate of objects under study

ii. Sampling method consume time and resource.

iii. Any summarised figure from population is known as statistics.

iv. We adopt sampling technique in our activities.

v. Population is a subset of sample
vi. An unbiased sample gives an accurate prediction of characteristics of an entire population.

vii. The standard deviation of sampling distribution of a statistic is known as standard error of the statistics.

viii. Standard error is used as a reliability measure.

ix. Faulty selection of sample contributes to sampling error.

x. Personal bias increases the non-sampling errors.

xi. Unbiased errors are cumulative in nature.

xii. Biased errors are also known as compensatory errors.

Answer:

1 i.=T, ii-F, iii-T, iv-F, vi-T, vii-T, viii-T, ix-T, x-F, xi-F, xii-F.
CHAPTER 8

Large Sample Theory

The difference between the hypothesized population parameter and the actual statistic is more often nightmares so large that we automatically reject our hypothesis nor so small that we just as quickly accept it. So in hypot etesting, as in most significant real-life decisions, clear-cut solutions are the exception not the rule.

Testing Hypothesis

Null and alternate hypothesis

In hypothesis testing, we must state the assumed or hypothesised value of the population parameter before we begin sampling. The assumption we wish to test is called the null hypothesis and is symbolised by $H_0$. The term 'null hypothesis' arises from earlier agricultural and medical applications of statistics.

In our sample results fail to support the null hypothesis, we must conclude that something else is true. Whenever we reject the hypothesis, the conclusion we do accept is called the alternative hypothesis and is symbolised $H_1$ ($H_{\text{sub-one}}$).

Selecting a Significance Level

There is no single standard or universal level of significance for testing hypotheses. In some instances, a 5% level of significance is used. In the published results of research papers, researchers often test hypothesis at the 1 per cent level of significance. Hence, it is possible to test a hypothesis at any level of significance. But remember that our choice of the minimum standard for an acceptable probability, or the significance level, is also the rest we assume of rejecting a null hypothesis when it is true. 5% level of significance implies we are ready to reject a true hypothesis in 5% of cases. If the significance level is high then we would rarely accept the null hypothesis when it is not true but, at the same time, often reject it when it is true. When testing the hypothesis we come across four possible situations. The table shows the four possible situations.
Table: Possible situations when testing a hypothesis.

<table>
<thead>
<tr>
<th>Test says</th>
<th>True</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accept</td>
<td>😊</td>
<td>😞</td>
</tr>
<tr>
<td>Reject</td>
<td>😞</td>
<td>😊</td>
</tr>
<tr>
<td>Hypothesis is</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type I error</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type II error</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The combinations are:

1. If the hypothesis is true, and the test result accepts it, then we have made a right decision.

2. If hypothesis is true, and the test results reject it, then we have made a wrong decision (Type I error). It is also known as Consumer's risk denoted by _____.

3. If hypothesis is false, and the test result accepts it, then we have made a wrong decision (Type II error). It is known as producer's risk, denoted by _____ is called power of the test.

4. Hypothesis is false, test result rejects it - we have made a right decision.

Preference of type I error

Suppose that making a Type I error (rejecting a null hypothesis when it is true) involves the time and trouble of reworking a batch of chemicals that should have been accepted.

Preference of type II error

Suppose on the other hand, that making a Type I error involves disassembling an entire engine at the factory, but making a Type II error involve relatively inexpensive warranty repairs by the dealers. Then the manufacturer is more likely to prefer a Type II error and will set lower significance level in its testing.

Determine appropriate distribution

After deciding what level of significance to use our next task in hypothesis testing is to determine the appropriate distribution. We have choice between the normal distribution, and the ‘t’ distribution.
Table: Conditions for using the normal and ‘t’ distributions in testing hypothesis about means

<table>
<thead>
<tr>
<th>Category</th>
<th>When the population Standard deviation is known</th>
<th>When the population Standard deviation is not known</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size 'n' is larger than 30</td>
<td>Normal distribution, z-table</td>
<td>Normal distribution, z-table</td>
</tr>
<tr>
<td>Sample size 'n' is 30 or less and we assume the population is normal or approximately so.</td>
<td>Normal distribution, z-table</td>
<td>‘t’ distribution, 't' table</td>
</tr>
</tbody>
</table>

Two-tailed test and one-tailed tests

Two tailed tests

A two tailed test of a hypothesis will reject the null hypothesis if the sample mean is significantly higher than or lower than the hypothesised population mean. Thus, in a two-tailed test, there are two rejection regions.

A two tailed test is appropriate when:

The null hypothesis is \( \mu = \mu_{Ho} \) (where \( \mu_{Ho} \) is some specified value)

the alternative hypothesis is \( \mu \neq \mu_{Ho} \).

One tailed

In general, a left tailed (lower tailed) test is used if the hypothesis are \( H_o: \mu = \mu_{Ho} \). In such a situation, it is sample evidence with the sample mean significantly below the hypothesised population mean that leads us to reject the null hypothesis in favour of the alternative hypothesis. Stated differently, the rejection region is in the lower tail (half tail) of the distribution of the sample mean, and that is why we call this a lower tailed test or (one tailed test).
Z - test

This test is used when the sample are large. We have seen that for large values of \( n \), the number of trials, almost all the distribution, \( e.g., \) binomial, poisson, etc., are very closely approximated by normal distribution. Thus in this case we apply the normal test, which is based upon the following fundamental property of the normal probability curve.

\[
Z = \frac{X - \mu}{\sigma} = \frac{X - E(X)}{\sqrt{V(X)}} \sim N(0,1)
\]

Thus from the normal probability tables, we have

\[
P(-3 \leq Z \leq 3) = 0.9973, \text{ i.e. } P(|Z| \leq 3) = 0.9973
\]

\[
P(|Z| > 3) = 1 - P(|Z| \leq 3) = 0.0027
\]

i.e in all probability we should expect a standard normal variate to lie between \( \pm 3 \).

also from the normal probability table we get

\[
P(-1.96 \leq Z \leq 1.96) = 0.95, \text{ i.e. } P(|Z| \leq 1.96) = 0.95
\]

\[
\Rightarrow P(|Z| > 1.96) = 1 - 0.95 = 0.05
\]

And

\[
P(|Z| > 2.58) = 0.99 \Rightarrow P(|Z| < 2.58) = 0.01
\]

Thus the significant values of \( Z \) at 5% and 1% level of significance for two tailed test are 1.96 and 2.58 respectively.

Thus the steps to be used in the normal test are as follows:

- Compute the test statistic \( Z \) under \( H_0 \).
- If \( |Z| > 3 \), \( H_0 \) is always rejected.
- If \( |Z| \leq 3 \), we test its significance at certain level of significance.
Example

A random samples 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 the size is 0.015 and deduce that the % of bad pineapples in the consignments almost certainly lies between 8.5 and 17.5.

Solution : Here we are given: n=500

X=numbers of bad pineapples in the sample=65

p= proportion of bad pineapples in the sample= 65/500 = 0.13 ⇒ q = 1− p = 0.87

since p the proportion of bad pineapple in the consignment is not known, we may take : P̂=p=0.13, Q̂=q=0.87.

S.E. of the proportion=√{P̂Q̂/ n} = √{0.13*0.87/500} = 0.015

Thus, the limits for the proportion of bad pineapples in the consignments are:

P̂± 3√{P̂Q̂/ n} = 0.130 ± 3×0.015 = 0.130 ± 0.45 = (0.085,0.175)

Hence the percentage of bad pineapples in the consignments almost certainly lies between 8.5 and 17.5.

Example. .

In a sample of 1000 people in Maharashtra, 540 are rice eaters and the rest are wheat eaters. Can we assume that both rice and wheat are equally popular in this state a 1% level of significance?

Solution. In the usual notations, we are given: n=1000

X = Number of rice eaters= 540

p̂ =sample proportion of rice eaters= X/n = 540/1000 = 0.54

Null Hypothesis, H₀ : both rice and wheat are equally popular in the state so that

p =population proportion of rice eaters in Maharashtra=0.5 ⇒ Q=1−p=0.5

Alternate Hypothesis,H₁ : p ≠ 0.5 (two tailed alternative)
Test statistic

\[ Z = \frac{p - P}{\sqrt{PQ/n}} \sim N(0, 1), \text{ (since n is large).} \]

Now

\[ Z = \frac{0.54 - 0.50}{0.5 \times 0.5 / 1000} = \frac{0.04}{0.0138} = 2.532 \]

Conclusion: The significant or critical value of Z at 1% level of significant for two tailed test is 2.58, since the computed Z = 2.532 is less than 2.58, it is not significant at 1% level of significance. Hence null hypothesis is accepted and we may conclude that rice and wheat are equally popular in Maharashtra.

**Self assessment questions**

1. For the following cases, specify which probability distribution to use in hypothesis testing:
   
   i. \( H_0: \mu = 27, H_1: \mu \neq 27, x = 33, \text{ sample } \sigma = 4, n = 25 \)
   
   ii. \( H_0: \mu = 98.6, H_1: \mu > 98.6, x = 99.1, \text{ sample } \sigma = 1.5, n = 50 \)
   
   iii. \( H_0: \mu = 3.5, H_1: \mu < 3.5, x = 2.8, \text{ sample } \sigma = 0.6, n = 18 \)
   
   iv. \( H_0: \mu = 382, H_1: \mu \neq 382, x = 363, \text{ sample } \sigma = 68, n = 12 \)
   
   v. \( H_0: \mu = 57, H_1: \mu > 57, x = 65, \text{ sample } \sigma = 12, n = 42 \)

2. State whether the following statements are 'True' or 'False'
   
   i. Null hypothesis state that there is significant difference between observed and hypothetical values.
   
   ii. 1% level of significance means we are ready to reject a true hypothesis in 99% cases.
   
   iii. If the Null hypothesis \( H_0: \) then it is two-tailed test.
   
   iv. If the calculated value of a statistic is less than tabulated value of the statistics, then \( H_0 \) is accepted.
   
   v. 1 - B is called fewer of the test.
   
   vi. If \( n_1 = 300, n_2 = 500, l = 50, 2 = 60, 1 = 10, 2 = 12 \) results of two samples taken from two cities A and B then we test for between means under different population.
   
   vii. If \( n > 30 \), then we do not apply z-test unless population S.D. is known.
1. Answers to Self Assessment

   i. Normal distribution
   ii. Normal distribution
   iii. 't' distribution DOF
   iv. Normal distribution
   v. Normal distribution

CHAPTER- 9

Theory of Estimation and Confidence Intervals

Everyone makes estimates. When you are ready to cross a street, you estimate the speed of any car that is approaching, the distance between you and that car, and your own speed. Having made these quick estimates, you decide whether to wait, walk or run. With the knowledge of inferential statistics, you can do the estimations about the population using the random samples which are drawn from the population.

Reasons for making estimates

All statisticians must make quick estimates. The outcome of these estimates can affect their organizations as seriously as the outcome of your decision whether to cross the street.

Making Statistical Inference

Statistical inference is based on estimation and hypothesis testing. In both estimation and hypothesis testing, we make inferences about characteristics of populations from information contained in sample.

Types of estimates

The following are two types of estimates about a population

i. Point estimates

ii. Interval estimates

Point estimate

Point estimates is a single number that is used to estimate an unknown population parameter. A point estimate is often insufficient, because it is either right or wrong. We don’t know how wrong it is. Therefore a point estimate is much more useful if it is accompanied by an estimate of the error that might be involved.

Interval estimate

Interval estimate is a range of values used in estimate a population parameter. It indicates the error in the following two ways:

i. By the extent of its range

ii. By the probability of the true population parameter lying within that range.
Example: The table below displays the results of samples of 35 boxes which contain bolts.

<table>
<thead>
<tr>
<th></th>
<th>101</th>
<th>103</th>
<th>112</th>
<th>102</th>
<th>98</th>
<th>97</th>
<th>93</th>
</tr>
</thead>
<tbody>
<tr>
<td>105</td>
<td>100</td>
<td>97</td>
<td>107</td>
<td>93</td>
<td>94</td>
<td>97</td>
<td></td>
</tr>
<tr>
<td>97</td>
<td>100</td>
<td>110</td>
<td>106</td>
<td>110</td>
<td>103</td>
<td>99</td>
<td></td>
</tr>
<tr>
<td>93</td>
<td>98</td>
<td>106</td>
<td>100</td>
<td>112</td>
<td>105</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>114</td>
<td>97</td>
<td>110</td>
<td>102</td>
<td>98</td>
<td>112</td>
<td>99</td>
<td></td>
</tr>
</tbody>
</table>

Considering the table, we have taken a sample of 35 boxes of bolts from a manufacturing line and have counted the bolts per box. We can arrive at the population mean that is, mean number of bolts by taking the mean for the 35 boxes we have sampled. This is calculated by adding all the bolts and dividing by the number of boxes.

$$\bar{x} = \frac{\sum x}{n} = \frac{3570}{35} = 102$$

Thus, using sample mean $\bar{x}$ as the estimator we have a point estimate of the population mean $\mu$.

Criteria of Good Estimator

Unbiasedness

Unbiasedness is a desirable property of a good estimator. The term unbiasedness refers to the fact that a sample mean is an unbiased estimator of a population mean because the mean of the sampling distribution of sample means taken from the same population is equal to the population mean itself.

Efficiency

Efficiency refers to the size of the standard error of the statistic. Let us compare two statistics from a sample of the same size and try to decide which one is the more efficient estimator. In this case, we would pick the statistic that has the smaller standard error.

Consistency

A statistic is a consistent estimator of a population parameter, if the sample size increases. It becomes almost certain that the value of the statistic comes very close to the
value of the population parameter, if an estimator is consistent, it becomes more reliable with large samples.

**Sufficiency**

An estimator is sufficient if it makes so much use of the information in the sample that no other estimator could extract from the sample any additional information about the population parameter being estimated.

**Interval Estimates and Confidence Intervals**

In using interval estimates, we are not confined to ±1, 2 and 3 standard errors; for example, ±1.64 standard errors include about 90 per cent of the area under the curve; it includes 0.4495 of the area on either side of the mean in a normal distribution. Similarly, ±2.58 standard error includes about 99 per cent of area, or 49.51 per cent on either side of the mean.

The probability that we associate with an interval estimate is called the confidence level.

This probability indicates how confident we are that the interval estimate will include the population parameter.

**Interval estimates of the mean of large samples**

If the samples are large, then we use the finite population multiplier to calculate the standard error. The standard error of the mean of finite population can be calculated as:

$$
\sigma_s = \frac{1}{\sqrt{n}} \times \frac{\sqrt{N-n}}{\sqrt{N-1}}
$$

and also the sample size ‘n’ is greater than five per cent of the population size ‘N’, i.e.

$$\frac{n}{N} > 0.05$$

**Interval estimates of the proportion of large samples**

Statisticians often use sample to estimate a proportion of occurrences in a population. For example, the government estimates, by a sampling procedure, the unemployment rate, or the proportion of unemployed people, in the country’s workforce.

We know that for a binomial distribution, the mean and the standard deviation are:

$$\text{Mean } \mu = np \\
\text{Standard deviation } \sigma = \sqrt{n \, p \, q}$$
Where,

\[ N = \text{number of trials} \]
\[ P = \text{probability of success} \]
\[ Q = \text{probability of failure} = 1 - p \]

Since we are taking the mean of the sample to be the mean of the population we actually mean that \( \mu = p \).

**Interval estimates using Student’s ‘t’ distribution**

So far, the sample sizes we were examining were all large than 30. This is not always the case. Questions like handling estimates where the normal distribution is not the appropriate sampling distribution are answered in this section.

**Degree of freedom**

There is a different ‘t’ distribution for each of the possible degrees of freedom.

**Self Assessment Questions**

Madhu, a frugal student, wants to buy a used bike. After randomly selecting 125 wanted advertisements, he found the average price of the bike to be Rs.3250 with a standard deviation of Rs.615. Establish an interval estimate for the average price of bike so that Madhu can be:

i. 68.3% certain that the population mean lies in this interval.
ii. 95.5% certain that the population mean lies in this interval.

Given the following confidence levels, express the lower and upper limits of the confidence interval for these levels in terms of \( \bar{x} \) and \( \sigma \) (use the normal distribution tables)

i. 54 per cent
ii. 75 per cent
iii. 94 per cent
iv. 98 per cent

From a population of 540, a sample of 60 individuals is taken. From this sample the mean is found to be 6.2 and the standard deviation to be 1.368.

i. Find the estimated standard error of the mean.
ii. Construct a 96% confidence interval of the mean.

For the following sample sizes and confidence levels, find the approximate ‘t’ values for constructing confidence intervals (use the ‘t’ table)

i. \( n = 28; 95\% \)
Answer to self assessment questions

The population standard deviation is given as:
\[ \sigma = 615; \quad n = 3250 \]

and standard error \( S_e \) is calculated as:
\[ \sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{615}{\sqrt{125}} = 55.01 \]

i. \( \bar{x} \pm 1 \sigma_x = 3250 \pm 55.01 = 3194.99 \) and 3305.01 to be 68.3% certain.

ii. 95.5% certain means \( \bar{x} \pm 2 \sigma_x = 3250 \pm 110.02 \) giving a range between 3139 and 3360.02.

iii. The required lower and upper class intervals are:
   i. \( \bar{x} \pm 0.74 \sigma_x \)  ii. \( \bar{x} \pm 1.15 \sigma_x \)  iii. \( \bar{x} \pm 1.88 \sigma_x \)  iv. \( \bar{x} \pm 2.33 \sigma_x \)

iv. i. \( \sigma_x = \frac{\sigma}{\sqrt{N}} x \frac{\sqrt{N-n}}{\sqrt{N-1}} as \frac{n}{N} > 0.05 \)

\[ \sigma_x = \frac{1.368 \times \sqrt{540 - 60}}{\sqrt{540 - 1}} = 0.167 \]

\[ \bar{x} \pm 2.05 \sigma_x = 6.2 \pm 2.05 (0.167) \] Hence, the LCL and UCL are 5.86 and 6.54, respectively.

i. 2.052  ii. 2.998  iii. 1.782  iv. 2.262
CHAPTER 10

Regression and Correlation

In this chapter, we shall deal with the dependence of factors or variables which take only numerical values. Very often, the interest to know the actual relationship between two or more variables. This problem is dealt with regression. On the other hand, we are often not interested to know the actual relationship but are only interesting to knowing the degree of relationship between two or more variables. This problem dealt with correlation analysis.

The concept of regression was first give by Sir Francis Galton (1822-1911) in a study of inheritance of stature in the human being. The use of regression technique becomes too common for a variety of problems. The relationship between variables, if it exists, may be linear or curvilinear. Linear relationship between two variables is represented by a straight line which is known as regression line. The line of average relationship is another name for a regression line. In the study of linear relationships between two variable $Y$ and $X$, suppose the variable $Y$ is such that it depends on $X$, then we call it the regression line of $Y$ on $X$. If $X$ depends on $Y$, it is called the regression of $X$ on $Y$. to find out the regression line, the observations $(x_i, y_i)$ on the variables $X$ and $Y$ are necessarily taken in pairs, on the unit my be people, animals plots, spare parts, plants or any other thing.

Scatter Diagram

When a regression equation is to be specified, n paired observations are plotted on the graph paper, setting the vertical scale for the dependent variable $Y$ and horizontal scale for the independent variable $X$. From the plotted points, it can easily be visualized, whether or not the plotted points lie in a straight line or appear to lie on a curve of a known type. As a matter of fact, scatter diagram may be considered as a basis of deciding the type of regression equation, suited for the relationship, between the two variables, $Y$ and $X$.

Least square method of fitting a regression line

The equation for a regression line of $Y$ on $X$ for the population is given as

$$Y = \alpha + \beta X + e$$

The above equation is also known as mathematical model for linear regression. The main difference between the cartesian equation of a line and a regression line is that a regression line is a probabilistic model which enables one to develop procedures for making inferences about the parameters $\alpha$ and $\beta$ of the model. In this model the expected
value of $Y$ is a linear function of $X$, but for fixed $X$, the variable $Y$ differs from its expected value by a random amount. As a special case the form $y = \alpha + \beta x$ is called the deterministic model.

The actual observed value of $y$ is a linear function of $x$. In this equation is the intercept which the line cuts on the axis of $Y$ and is the slope of the line. $\beta$ is also called the regression coefficient and is defined as, “$\beta$ is the measure of change in the dependent variable ($Y$) corresponding to a unit change in the independent variable ($X$). $\beta$ can take any real value within the range $-\infty$ to $\infty$.

**Regression coefficient**

When a regression is a linear, then the regression coefficient is given by the slope of the regression line.

a. The geometric mean of regression coefficients gives the correlation coefficient.

$$b_{yx}.b_{xy} = r^2$$

$$\sqrt{b_{yx}.b_{xy}} = 1$$

b. The product of regression coefficients is always less than “1”

$$b_{yx}.b_{xy} \leq 1$$

c. If $b_{yx}$ is negative, then $b_{xy}$ is also negative and ‘r’ is negative

d. They can also be expressed as

$$b_{yx} = r \sigma_y/\sigma_x and b_{xy} = r \sigma_x/\sigma_y$$

e. It is an absolute measures

**Properties of regression coefficient**

It can take any value between $-\infty$ to $\infty$. Its sign is same as that of $s_{XY}$ i.e. $\text{cov}(X, Y)$

Further, if the whole population has been studied. $i$ will vary from 1 to N for all the N units of the population. In this situation we get population regression coefficient $\beta$ directly, for which the formula is:

$$\beta = \frac{\sigma_{XY}}{\sigma_x^2}$$

In case of population regression coefficient $\beta_y$ of $Y$ on $X$, which can elaborately be specified as $\beta_{YX}$ we can express it as,

$$\beta_{YX} = \rho \frac{\sigma_y}{\sigma_x}$$

Where, $\rho$ is the population correlation coefficient between $X$ and $Y$. 

---

*Teaching Manual on “Statistical Methods for Applied Sciences” - Ravi R. Saxena, M.L. Lakhera and Roshan Bhardwaj, Department of Agricultural Statistics and SS (E), IGKV, Raipur – 492 012 (Chhattisgarh)*
As \( a \) and \( b \) are the estimated values of \( \alpha \) and \( \beta \), respectively, the equation of the estimated regression line is

\[
\hat{Y} = a + bX
\]

Where, the hat (\(^{\wedge}\)) over \( Y \) indicates that \( Y \) is an estimated value. Substituting the value the line of best fit is,

\[
\hat{Y} = (\bar{Y} - b\bar{X}) + bx
\]

\[
(\hat{Y} - \bar{Y}) = b(X - \bar{X})
\]

**Prediction equation:**

The regression line is also known as prediction equation. Once the constant \( a \) and \( b \) are calculated, there remains two unknown variables in the regression equation viz., \( Y \) and \( X \). Moreover, we know \( Y \) depend on \( X \) in the case of regression equation of \( Y \) on \( X \). Under the presumption that the trend of change in \( Y \) corresponding to \( X \) remains the same, the value of \( Y \) can be estimated for any value of \( X \). But such a presumption rarely holds good for a very wide range of \( X \) values.

Example: The table below gives the data regarding fertilizer consumption and production (taking 1960 = 100) from 1951 to 1970.

<table>
<thead>
<tr>
<th>Year</th>
<th>Fertilizer consumption (( X ))</th>
<th>Production (( Y ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1951</td>
<td>36.6</td>
<td>54.8</td>
</tr>
<tr>
<td>1952</td>
<td>39.5</td>
<td>57.2</td>
</tr>
<tr>
<td>1953</td>
<td>43.4</td>
<td>58.1</td>
</tr>
<tr>
<td>1954</td>
<td>47.6</td>
<td>63.4</td>
</tr>
<tr>
<td>1955</td>
<td>53.4</td>
<td>72.5</td>
</tr>
<tr>
<td>1956</td>
<td>58.5</td>
<td>78.4</td>
</tr>
<tr>
<td>1957</td>
<td>66.1</td>
<td>82.7</td>
</tr>
<tr>
<td>1958</td>
<td>74.9</td>
<td>84.4</td>
</tr>
<tr>
<td>1959</td>
<td>87.1</td>
<td>90.3</td>
</tr>
<tr>
<td>1960</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>1961</td>
<td>115.1</td>
<td>109.2</td>
</tr>
<tr>
<td>1962</td>
<td>131.7</td>
<td>119.8</td>
</tr>
<tr>
<td>1963</td>
<td>150.0</td>
<td>129.7</td>
</tr>
<tr>
<td>1964</td>
<td>162.6</td>
<td>140.8</td>
</tr>
<tr>
<td>1965</td>
<td>176.3</td>
<td>153.8</td>
</tr>
</tbody>
</table>
It is known that the production \((Y)\) depends on the consumption of fertilizer \((X)\) and the relation between the two variables is linear. Hence, a regression line can be fitted to the bivariate data by calculating the values of \(a\) and \(b\). First make the following calculations.

\[
\begin{align*}
\sum x_i &= (36.6 + 39.5 + \ldots + 271.4) = 2503.3 \\
\sum y_i &= (54.8 + 57.2 + \ldots + 184.3) = 2223.5 \\
\sum x_i^2 &= (36.6^2 + 39.5^2 + \ldots + 271.4^2) = 425211.85 \\
\end{align*}
\]

\[
\therefore \quad \bar{x} = \frac{2503.3}{20} = 125.17 \\
\bar{y} = \frac{2223.5}{20} = 111.81
\]

Also \(n = 20\)

\[
\sum x_i y_i = (36.6 \times 54.8 + 39.5 \times 57.2 + \ldots + 271.4 \times 184.3)
\]
\[
= 339769.87
\]

Using the formula:

\[
b = \frac{339769.87 - \frac{1}{20} (2503.3)(2223.5)}{425211.85 - \frac{1}{20} (2503.3)^2}
\]

\[
= \frac{61465.49}{111886.31} = 0.55
\]

And

\[
a = 111.18 - 0.55 \times 125.17 = 42.34
\]

Hence, the estimated equation of the regression line is:

\[\hat{Y} = 42.34 + 0.55X\]

Given the value of \(X = 150\), we can estimate the value of \(Y\) from the estimated equation, that is

\[\hat{Y} = 42.34 + 0.55 \times 150 = 124.84\]

If we see that data, we find that the actual value of \(Y\) for \(X = 150\) is 129.7. The difference between the actual and estimated value is not much. Hence, the line seems to
be good fit. Again if we want the projection for the increased consumption of electricity, i.e. X=350, the estimated value of Y is,

\[ \hat{Y} = 42.34 + 0.55 \times 350 = 234.84 \]

From this we infer that if the consumption of fertilizer rises to the level of 350, production will rise to the level of 234.84.

**Regression line of X on Y:**

Often we come across situations in which two variables (X and Y) are such that not only Y depends on X but X also depends on Y. For example, the heights and weights of people are two variables where heights of people depends on weights and weights depends on heights. In such a case we can find not only the regression line of Y on X but also of X on Y. Suppose the regression line of X on Y is,

\[ X = \alpha_1 + \beta_1 Y + e_1 \]

The parameters \( \alpha_1 \) and \( \beta_1 \) can be estimated in the same way as \( \alpha \) and \( \beta \). Instead of repeating the derivation, it will be worthwhile to write directly the estimated values of \( \alpha_1 \) and \( \beta_1 \), say \( a_1 \) and \( b_1 \) (by interchanging the variable Y by X and X by Y in the formulae for a and b respectively. Thus, the estimates are

\[
\begin{align*}
    a_1 &= (\bar{x} - b_1 \bar{y}) \\
    b_1 &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (y_i - \bar{y})^2} \\
        &= \frac{\sum x_i y_i - (\sum x_i)(\sum y_i)/n}{\sum y_i^2 - (\sum y_i)^2/n} \\
        &= \frac{S_{XY}}{S_Y^2}
\end{align*}
\]

We can express \( b_1 \), which is the estimated regression coefficient of X on Y and symbolically denoted as \( b_{XY} \). In terms of \( r, s_X, s_Y \), we can write.

\[
b_{XY} = \frac{S_X}{S_Y}
\]

Where r is the sample correlation coefficient between X and Y.

The equation of the estimated regression line of X on Y is,

\[
\left( X - \bar{x} \right) = b_1 \left( Y - \bar{y} \right)
\]

Also the population coefficient of X on Y may be given as follows:

\[
\beta_1 = \frac{\sigma_{XY}}{\sigma_Y^2}
\]
The population regression coefficient $\beta_1$ of $X$ on $Y$, which is often symbolized as $\beta_{XY}$, can be expressed as,

$$\beta_{XY} = \rho \frac{\sigma_X}{\sigma_Y}$$

Where is the population correlation between $X$ and $Y$.

It is trivial to prove that the two regression lines given intersect each other at a point having coordinates $(\bar{x}, \bar{y})$, i.e. at the mean of two variables.

Example: Using the data and partial calculations of example, we fit in the regression line.

$$\left( X - \bar{x} \right) = b \left( Y - \bar{y} \right)$$

First we calculate $\sum_i y_i^2 = (54.8^2 + 57.2^2 + ... + 184.3^2)$

$$= 281790.03$$

Now from

$$b_i = \frac{61465.49}{281790.03 - \frac{1}{20}(2223.5)^2}$$

$$= \frac{61465.49}{34592.42}$$

$$= 1.78$$

Hence, the required equation for the regression line is,

$$\left( \hat{X} - 125.17 \right) = .78(Y - 111.18)$$

or

$$\hat{X} = .78Y - 72.73$$

The value of $X$ can be estimated for any give value of $Y$, in the manner followed in example.

**Estimation of Parameters**

Test of significance of Regression Parameters: The estimators of $\alpha$ and $\beta$ have already been calculated based on $n$ paired observations. The experimenter is interested to test whether the parameters $\alpha$ and $\beta$, involved in the regression line, are of practical relevance or not. To do so, we have to test the hypotheses $\beta = 0$ and $\alpha = 0$. These hypotheses can be tested provided we know the distribution of $b$ and $a$. The estimators $a$ and $b$ are random variables such that:

$$b \sim N\left( \beta, \sigma_b^2 \right)$$

and

$$a \sim N\left( \alpha, \sigma_a^2 \right)$$
Confidence Limits of Regression Parameters:

The (1 - \(\alpha\)) per cent confidence limit is given by.

Confidence limits for \(\beta_{YX}\) are,

\[ b \pm S_b \ t_{\alpha,(n-2)} \]

Where, \(t_{\alpha,(n-2)}\) is the table value for two tailed t-test at \(\alpha\) level of significance and for (n-2) degree of freedom, \(b\) and \(S_b\) are as given earlier.

Again (1 - \(\alpha\)) per cent confidence limits for \(\alpha\), the intercept, are,

\[ a \pm S_a \ t_{\alpha,(n-2)} \]

Where, \(a\) and \(S_a\) are as given earlier, \(t_{\alpha,(n-2)}\) has been explained just before.

Example : We give the tests of significance for the data given in example

Test of significance of regression coefficient amounts to testing of hypothesis,

\( H_0 : \beta_{YX} = 0 \) vs. \( H_1 : \beta_{YX} \neq 0 \)

To test \(H_0\), we make use of the test statistic,

\[ t_{n-2} = \frac{b}{S_b} \]

Residuals: The difference between the observed value of the dependent variable (\(y\)) and the predicted value (\(\hat{y}\)) is called the residual (\(e\)). Each data point has one residual.

\[ e = y - \hat{y} \]

Both the sum and the mean of the residuals are equal to zero. That is, \(\Sigma e = 0\) and \(e = 0\).

Residual Plots

A residual plot is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.

Below the table on the left inputs and outputs for a simple linear regression analysis, and the chart on the right displays the residual (\(e\)) and independent variable (\(X\)) as a residual plot.
The residual plot shows a fairly random pattern - the first residual is positive, the next two are negative, the fourth is positive, and the last residual is negative. This random pattern indicates that a linear model provides a decent fit to the data.

Below, the residual plots show three typical patterns. The first plot shows a random pattern, indicating a good fit for a linear model. The other plot patterns are non-random (U-shaped and inverted U), suggesting a better fit for a non-linear model.

we will work on a problem, where the residual plot shows a non-random pattern. And we will show how to "transform" the data to use a linear model with nonlinear data.

Test Your Understanding of This Lesson

In the context of regression analysis, which of the following statements are true?

I. When the sum of the residuals is greater than zero, the data set is nonlinear.
II. A random pattern of residuals supports a linear model.
III. A random pattern of residuals supports a non-linear model.

(A) I only  (B) II only  (C) III only  (D) I and II  (E) I and III

Solution

The correct answer is (B). A random pattern of residuals supports a linear model; a non-random pattern supports a non-linear model. The sum of the residuals is always zero, whether the data set is linear or nonlinear.
Correlation

Regression technique provides the actual relationship between two or more variables. But scientists are not always interested in this linear or curvilinear relationship. Often the interest lies only in knowing the extent of interdependence between two or more variables. In this situation, correlation methods serve our purpose. If the two variables, say X and Y, are linearly related, they are said to be correlated. The correlation between two variables is accompanied by the proportionate change in the other variable.

The mathematical measure of correlation was given by the statistician Karl Pearson in 1986 in the form of correlation coefficient, by Professor Neiswanger as he wrote, “correlation analysis contributes to the understanding of economic behavior, aids in locating the critically important variable on which other depend, may reveal to the economist the connections by which disturbances spread and suggest to him the paths through which stabilising forces may become effective.

**Determination of Correlation by Graphic method:** It has already been stated, while explaining the regression, that scatter diagram gives a fair idea of relationship between two variables, X and Y. The population correlation coefficient is denoted by ρ (rho).

**Types of Correlation**

There are important types of correlation. They are:
1. Positive correlation.
2. Perfectly Positive Correlation
3. Negative correlation
4. Perfectly negative Correlation
5. Linear correlation.

**(1) Positive Correlation.**

When two variables move in the same direction then the correlation between these two variables is said to be Positive Correlation. When the value of one variable increases, the value of other value also increases at the same time.
Some examples of series of Positive correlation are:

- i) Heights and weights;
- (ii) Household income and expenditure;
- (iii) Price and supply of commodities;
- (iv) Amount of rainfall and yield of crops.
The correlation between HBZ-SI mRNA and HTLV-1 proviral loads. The positive correlation between the expression level of HBZ-SI mRNA (Y-axis) and the proviral load (X-axis) equivalent to the infected cell number (r = 0.483, P < 0.05), indicating that an adjusted HBZ-SI value is indispensable to evaluate the expression level in heterogeneous samples with a mixture of infected and uninfected cells.

Negative Correlation.

In this type of correlation, the two variables move in the opposite direction.

When the value of a variable increases, the value of the other variable decreases.

For example:-, The relationship between price and demand.

Perfect Positive Correlation.

When there is a change in one variable, and if there is equal proportion of change in the other variable say Y in the same direction, then these two variables are said to have a
Perfectly Negative Correlation:

*Between two variables X and Y, if the change in X causes the same amount of change in Y in equal proportion but in opposite direction, then this correlation is called as Perfectly Negative Correlation.*

Zero Correlation:

*When the two variables are independent and the change in one variable has no effect in other variable, then the correlation between these two variables is known as Zero Correlation.*
Linear Correlation:

The correlation between two variables is said to be linear if the change of one unit in one variable result in the corresponding change in the other variable over the entire range of values.

For example consider the following data.

<table>
<thead>
<tr>
<th>X</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>7</td>
<td>13</td>
<td>19</td>
<td>25</td>
<td>31</td>
</tr>
</tbody>
</table>

Thus, for a unit change in the value of x, there is a constant change in the corresponding values of y and the above data can be expressed by the relation \( y = 3x + 1 \)

In general two variables \( x \) and \( y \) are said to be linearly related, if there exists a relationship of the form  \( y = a + bx \)

where ‘a’ and ‘b’ are real numbers. This is nothing but a straight line when plotted on a graph sheet with different values of x and y and for constant values of a and b. Such relations generally occur in physical sciences but are rarely encountered in economic and social sciences.

If the quantum of change in one variable has a ratio of change in the quantum of change in the other variable then it is known as Linear correlation.

Nonlinear correlation

Any correlation in which the rates of change of the variables is not constant.
Correlation coefficient

The correlation between the two variables is termed as simple correlation and is a general measure of Karl Pearson coefficient of correlation. The estimated value of population correlation \( \rho \) between two variables \( X \) and \( Y \) is denoted by \( r \). Here we shall give all the formula for \( r \).

Sometimes the suffix \( XY \) is added to \( r \), i.e. \( r_{XY} \), to connote that it is the correlation coefficient between the variable \( X \) and \( Y \). Theoretically, the sample correlation coefficient is given as

\[
r_{XY} = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \cdot \text{var}(Y)}} = \frac{s_{XY}}{s_X \cdot s_Y}
\]

Where \( s_X \) and \( s_Y \) are the sample standard deviations of variables \( X \) and \( Y \) respectively, and \( s_{XY} \) is the estimated covariance between \( X \) and \( Y \). If we have \( n \) pairs of sample observations, \( (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \), then the correlation coefficient.

\[
r = \frac{\frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})}{\left\{ \frac{1}{n-1} \sum (x_i - \bar{x})^2 \right\}^{\frac{1}{2}} \left\{ \frac{1}{n-1} \sum (y_i - \bar{y})^2 \right\}^{\frac{1}{2}}}
\]

for \( i = 1, 2, \ldots, n \)

\[
= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}
\]
\[
\begin{align*}
\text{for } i = 1, 2, \ldots, n \\
\text{If we take the deviations from mean and denote} \\
x_i - \bar{x} = u_i \text{ and } y_i - \bar{y} = v_i, \\
\text{Then the coefficient of correlation} \\
r = \frac{\sum u_i v_i}{\sqrt{\sum u_i^2 \sum v_i^2}} \\
\text{The correlation coefficient } r \text{ is a pure number i.e. } r \text{ is independent of the units in which } X \text{ and } y \text{ are measured.}
\end{align*}
\]

**Assumptions about Correlation Coefficient:**

There are three assumption made in giving the correlation coefficient by the above formulae. They are:

1. The random variables $X$ and $Y$ are distributed normally.
2. The variable $X$ and $y$ are linearly related.
3. There is a cause and effect relationship between factors affecting the values of $X$ and $Y$ in the series of data.

**Limits of Correlation Coefficient:** Formula gives,

\[
r = \frac{\sum u_i v_i}{\sqrt{\sum u_i^2 \sum v_i^2}}
\]

We know from the Schwartz inequality that if $u_i$ and $v_i$ are real quantities for all $i = 1, 2, \ldots, n$, then

\[
(\sum u_i v_i)^2 \leq (\sum u_i^2)(\sum v_i^2)
\]

The sign of equality holds if and only if

\[
\frac{u_1}{v_1} = \frac{u_2}{v_2} = \ldots = \frac{u_n}{v_n}
\]

Thus using (13.39), we can write that

\[
r^2 \leq 1
\]

Which implies that $-1 \leq r \leq 1$. 
**Example**: The age in years of fourteen young couples is given below:

<table>
<thead>
<tr>
<th>Husband (X)</th>
<th>21</th>
<th>25</th>
<th>26</th>
<th>24</th>
<th>22</th>
<th>30</th>
<th>19</th>
<th>24</th>
<th>28</th>
<th>32</th>
<th>31</th>
<th>29</th>
<th>21</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wife (Y)</td>
<td>19</td>
<td>20</td>
<td>24</td>
<td>21</td>
<td>21</td>
<td>24</td>
<td>18</td>
<td>22</td>
<td>19</td>
<td>30</td>
<td>27</td>
<td>26</td>
<td>19</td>
<td>18</td>
</tr>
</tbody>
</table>

To know the extent of relationship between the age of husbands and wives, we calculate the coefficient of correlation $r$ from the given data. For the given variate values,

\[ N = 14, \sum x_i = 350, \bar{x} = 25, \sum y_i = 308, \bar{y} = 25, \]

Since means are whole numbers, it is convenient to make use of the formula. To do the calculations systematically, prepare the following table.

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>$x - \bar{x}$</th>
<th>$y - \bar{y}$</th>
<th>$(x - \bar{x})(y - \bar{y})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>19</td>
<td>-1</td>
<td>-3</td>
<td>12</td>
</tr>
<tr>
<td>25</td>
<td>20</td>
<td>0</td>
<td>-2</td>
<td>0</td>
</tr>
<tr>
<td>24</td>
<td>24</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>26</td>
<td>21</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>22</td>
<td>21</td>
<td>-3</td>
<td>-1</td>
<td>3</td>
</tr>
<tr>
<td>30</td>
<td>24</td>
<td>5</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>19</td>
<td>18</td>
<td>-6</td>
<td>-4</td>
<td>24</td>
</tr>
<tr>
<td>24</td>
<td>22</td>
<td>-1</td>
<td>0</td>
<td>24</td>
</tr>
<tr>
<td>28</td>
<td>19</td>
<td>3</td>
<td>-3</td>
<td>-9</td>
</tr>
<tr>
<td>32</td>
<td>30</td>
<td>7</td>
<td>8</td>
<td>56</td>
</tr>
<tr>
<td>31</td>
<td>27</td>
<td>6</td>
<td>5</td>
<td>30</td>
</tr>
<tr>
<td>29</td>
<td>26</td>
<td>4</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>21</td>
<td>19</td>
<td>-4</td>
<td>-3</td>
<td>12</td>
</tr>
<tr>
<td>18</td>
<td>18</td>
<td>7</td>
<td>-4</td>
<td>28</td>
</tr>
</tbody>
</table>

From the above table we have:
\[ \sum(x - \bar{x})^2 = 264, \sum(y - \bar{y})^2 = 178, \sum(x - \bar{x})(y - \bar{y}) = 185 \]

Putting these values in the formula (13.37.1 we get:

\[ r = \frac{185}{\sqrt{264 \times 178}} = \frac{185}{216.78} = 0.85 \]

The coefficient of correlation between the age of husband and that of wife is 0.85 which is close to 1. Hence, it can be said that there is a high degree of relationship between the age of husbands and wives.

Example:

The birth rate and death rate per thousand persons in India from 2000 to 2010 were as follows:

<table>
<thead>
<tr>
<th>Birth rate (X)</th>
<th>16.5</th>
<th>15.8</th>
<th>15.2</th>
<th>14.3</th>
<th>13.6</th>
<th>12.9</th>
<th>12.3</th>
<th>11.7</th>
<th>11.5</th>
<th>11.3</th>
<th>11.3</th>
<th>11.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Death rate (Y)</td>
<td>9.3</td>
<td>9.1</td>
<td>8.2</td>
<td>8.9</td>
<td>8.9</td>
<td>8.5</td>
<td>9.7</td>
<td>9.0</td>
<td>8.7</td>
<td>9.1</td>
<td>9.0</td>
<td>9.2</td>
</tr>
</tbody>
</table>

The correlation between the birth rate and the death rate can be calculated in the following manner.

<table>
<thead>
<tr>
<th>Obs. No.</th>
<th>x</th>
<th>y</th>
<th>x^2</th>
<th>y^2</th>
<th>xy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17.1</td>
<td>9.3</td>
<td>292.41</td>
<td>86.49</td>
<td>159.03</td>
</tr>
<tr>
<td>2</td>
<td>16.5</td>
<td>9.3</td>
<td>272.25</td>
<td>86.49</td>
<td>153.45</td>
</tr>
<tr>
<td>3</td>
<td>15.8</td>
<td>9.1</td>
<td>249.64</td>
<td>82.81</td>
<td>143.78</td>
</tr>
<tr>
<td>4</td>
<td>15.2</td>
<td>8.2</td>
<td>231.04</td>
<td>67.24</td>
<td>124.64</td>
</tr>
<tr>
<td>5</td>
<td>14.3</td>
<td>8.9</td>
<td>204.49</td>
<td>79.21</td>
<td>127.27</td>
</tr>
<tr>
<td>6</td>
<td>13.6</td>
<td>8.9</td>
<td>184.96</td>
<td>79.21</td>
<td>121.04</td>
</tr>
<tr>
<td>7</td>
<td>12.9</td>
<td>8.5</td>
<td>166.41</td>
<td>72.25</td>
<td>109.65</td>
</tr>
<tr>
<td>8</td>
<td>12.3</td>
<td>9.7</td>
<td>151.29</td>
<td>94.09</td>
<td>119.31</td>
</tr>
<tr>
<td>9</td>
<td>11.7</td>
<td>9.0</td>
<td>136.89</td>
<td>81.00</td>
<td>105.30</td>
</tr>
<tr>
<td>10</td>
<td>11.5</td>
<td>8.7</td>
<td>132.25</td>
<td>75.69</td>
<td>100.05</td>
</tr>
<tr>
<td>11</td>
<td>11.3</td>
<td>9.1</td>
<td>127.69</td>
<td>82.81</td>
<td>102.83</td>
</tr>
</tbody>
</table>
There are 13 pairs of observations, hence $n=13$. Now we make use of formula because of means of $x$ and $y$ values are not exact.

$$r = \frac{1574.77 - (175.1)(116.9)/13}{\sqrt{2411.57 - (175.1^2)/13} \left(1052.93 - (116.9^2)/13\right)}$$

$$= \frac{0.22}{\sqrt{53.11 \times 1.73}} = \frac{0.22}{9.58} = 0.02$$

The coefficient of correlation between the birth rate and the death rate is 0.02, which is very close to zero. Hence, it can be inferred they are not linearly related to each other.

**Test of significance of Correlation coefficient**

The random sample (s) is drawn from the population or universe sample under consideration. Whatever conclusions are derived or deduced from the sample values, are meant to draw inferences about the parent population. The estimates are not unique and hence a sort of confirmation is sought by way of test of significance, for the validity of inferences drawn from the sample about the population. Of course, these results are subjected to certain probability of wrong decision which is covered by the level of significance.

The test of significance of correlation coefficient means to test the hypothesis, whether or not the correlation coefficient is zero in the population i.e. we test,

$$H_0 : \rho = 0 \text{ vs. } H_1 : \rho \neq 0$$

The test of statistic for testing $H_0$ is

$$t_{n-2} = \frac{r}{S_r}$$

Where $r$ is the estimated value of $\rho$ based on the n paired observation and $S_r$ is the standard error of $r$. Suffix (n-2) denotes the degree of freedom of t. Also

$$S_r = \sqrt{\frac{1 - r^2}{n-2}}$$
\[ t_{n-2} = \frac{r \sqrt{n-2}}{\sqrt{1-r^2}} \]

Example: Given that 
\( r = 0.9958 \) and \( n = 9 \)

To test the significance of the population correlation coefficient, we test 
\( H_0: \rho = 0 \) vs. \( H_1: \rho \neq 0 \)

By statistic (13.45.1) Hence, 
\[ t = \frac{0.9958 \sqrt{9-2}}{\sqrt{1-(0.9958)^2}} \]
\[ = \frac{0.9958 \sqrt{7}}{\sqrt{0.0084^2}} = \frac{2.63}{0.092} = 28.59 \]

The table value of \( t \) at \( \alpha = 0.05 \) and 7 d.f. is 2.365. Since the calculated value of \( t \) is greater than the table value of \( t \), we reject \( H_0 \). It means that there is a significant correlation between the area sown and the production.

**Relationship between correlation coefficient and regression coefficients**

It has been said that simple correlation is expressed only when there exists a linear relation between two variables. Hence, it should be possible to establish the mathematical relationship between the correlation coefficient and the two regression coefficients namely \( b_{YX} \) and \( b_{XY} \), we known

\[ r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}} \]

For \( i = 1, 2, \ldots, n \)

Or \[ r^2 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})^2}{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2} \]

Also, \( b_{YX} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \)

\[ b_{XY} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (y_i - \bar{y})^2} \]

\[ \therefore b_{YX} b_{XY} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})^2}{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2} \]

Relation shows that the correlation coefficient is the geometric mean of the two regression coefficients. The sign of \( r \) will be the same as that of either \( b_{YX} \) or \( b_{XY} \).
Again \[ b_{XY} = \frac{S_{XY}}{S_X^2} \]

And \[ r = \frac{S_{XY}}{S_X S_Y} \]

\[ S_{XY} = r S_X S_Y \]

Substituting for \( S_{XY} \) we get,

\[ b_{XY} = r \frac{S_Y}{S_X} \]

Similarly, \[ b_{XY} = r \frac{S_X}{S_Y} \]

**Rank Correlation**

Many characters are expressed in comparative terms such as beauty, smartness, temperament, etc. In such cases the units are ranked pertaining to that particular character instead of taking measurements on them. Sometimes, the units are also ranked according to their quantitative measure. In these types of studies, two situations arise (i) the same set of units is ranked according to two characters A and B, (ii) two judges give ranks to the same set of unit independently, pertaining to one character only. In both these situations, we get paired ranks for a set of units. The psychologist, Charles Edward Spearman (1906) developed a formula for correlation coefficient, which is known as rank correlation or Spearman’s correlation. It is denoted by \( r_S \) Suffix S to r is a connotation for Spearman, the name of the inventor. The formula for \( r_S \) is derived as the ratio of covariance to the product of standard deviation of two series of ranks. Here the deviation is omitted and thus the formula for rank correlation is:

\[ r_S = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)} \]

Where, \( i = 1, 2, \ldots, n \)

\( d_i \) is the difference between the rank of the \( i^{th} \) units and n is the total number of units.

The value of \( r_S \) lies between -1 and 1.

Example :

Two judges gave the following ranks (from the highest to the lowest) to eleven girls who contested in a beauty competition. Whether or not, there is an agreement between the independent rankings of the two judges, can be ascertained only by finding out the rank correlation between the ranks awarded by two judges.
Following the usual procedure, the rank correlation is calculated.

\[
d_1 = 1, 0, -2, 1, -2, 1, 5, -4, -1, 9, -2 \quad \text{Total}
\]

\[
d_2 = 1, 0, 4, 1, 4, 1, 25, 16, 1, 9, 4 \quad 66
\]

Here \( \sum d_i^2 = 66 \) and \( n = 11 \)

From formula

\[
r_s = 1 - \frac{6 \times 66}{11 \times 120} = 1 - 0.30 = 0.70
\]

The value of rank correlation \( r_s = 0.70 \), which is quite high. Hence, it can be concluded that there is an agreement between judges with regard to the beauty of the girls.

Example:

The ranks of 12 students according to their marks in Mathematics and Statistics were as follows:

<table>
<thead>
<tr>
<th>Student No.</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>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematics</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>8</td>
<td>11</td>
<td>12</td>
<td>4</td>
<td>3</td>
<td>9</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>Statistics</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>7</td>
<td>6</td>
<td>9</td>
<td>10</td>
<td>5</td>
<td>1</td>
<td>11</td>
<td>8</td>
<td>12</td>
</tr>
</tbody>
</table>

The interest lies to know whether or not students who are good in Mathematics also excel in statistics and vice versa. This objective can be met out by finding out the rank correlation coefficient. For this the calculations are:

\[
d_1 = 1, -1, -1, -1, 2, 2, 2, -1, 2, -2, -1, -2 \quad 0
\]

\[
d_2 = 1, 1, 1, 1, 4, 4, 4, 1, 4, 4, 1, 4 \quad 30
\]
Here \[ \sum_i d_i^2 = 30 \] and \( n = 12 \)

\[ r_s = 1 - \frac{6 \times 30}{12 \times 143} = 1 - 0.105 = 0.895 \]

The correlation between the ranks of marks in the two subjects is very high. From this, it is inferred that the students who are good in Mathematics are also good in Statistics.

**Coefficient of Determination:**

The coefficient of determination, which is given as \( r^2 \), explains to what extent the variation of dependent variable \( Y \) is being expressed by the independent variable \( X \).

\[ r^2 = \frac{\text{Variance explained}}{\text{Total variance}} \]

More the value of \( r^2 \), better it is. A high value clearly shows that a good linear relation exists between the two variables. Obviously, if \( r=1 \), then \( r^2=1 \), which is an indicator of perfect relationship between the two variables. Also the quantity \( 1-r^2 \) is called the coefficient of non-determination or coefficient of alienation \( (1-r^2) \) is thus a measure of deviation from perfect linear relationship.

**Multiple and partial correlation**

While fitting a regression equation, the interest lies to know how far this equation serves our purpose. That is we want to confirm whether the equation is a good fit or not. In other words to what extent do the predictors explain the predictant \( Y \)? One way of describing the relative goodness of fit is the coefficient of determination \( R^2 \). \( R^2 \) tells what part of the total variation \( \sum_i (y_i - \bar{y})^2 \) in \( Y \) is explained by \( X \)-variables. Formula for \( R^2 \) is,

\[ R^2 = \frac{\sum_i v_i^2}{\sum_i v_i^2} \]

Where, \( i = 1, 2, \ldots, n \).

\[ \sum_i v_i^2 = \sum_i (y_i - \bar{y})^2 = \sum_i y_i^2 - (\sum_i y_i)^2 / n \]

**Definitions**

i. Multiple correlation coefficient is a measure of linear association of a variable \( Y \) (say) with \( X \)-variables

ii. Multiple correlation coefficient is the simple correlation between \( Y \) and \( \hat{Y} \) where \( \hat{Y} = \bar{Y} + b_1X_1 + b_2X_2 + \ldots + b_kX_k \)

The range of \( R \) is also from 0 to 1.
Multiple correlation

Three or more variables are involved in multiple correlations. The dependent variable is denoted by \( X_1 \) and other variables are denoted by \( X_2, X_3 \) and so on. Gupta SP has expressed that the “coefficient of multiple linear correlation is represented by \( R_1 \) and it is common to add subscripts designating the variables involved. Thus, \( R_{1.234} \) would represent the coefficient of multiple linear correlations between \( X_1 \) on the one hand \( X_2, X_3 \) and \( X_4 \) on the other. The subscript of the dependent variable is always to the left of the point.

The coefficient of multiple correlations for \( r_{12}, r_{13} \) and \( r_{23} \) can be expressed as

\[
R_{1.23} = \sqrt{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}/1 - r_{23}^2}
\]

\[
R_{2.13} = \sqrt{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}/1 - r_{13}^2}
\]

\[
R_{3.12} = \sqrt{r_{13}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}/1 - r_{12}^2}
\]

Coefficient of multiple regression \( R_{1.23} \) is the same as \( R_{1.32} \).

A coefficient of multiple correlation lies between ‘0’ and ‘1’. If the coefficient of multiple correlations is ‘1’, it shows that the correlation is perfect. If it is ‘0’, it shows that there is no linear relationship between the variables. The coefficients of multiple correlations are always positive in sign and range from ‘0’ to ‘+1’. Coefficient of multiple determinations can be obtained by squaring \( R_{1.23} \).

Alternative formula for computing \( R_{1.23} \) is:

\[
R_{1.23} = \sqrt{r_{12}^2 + r_{13}^2 (1 - r_{12}^2)}
\]

\[
R_{1.23}^2 = r_{12}^2 + r_{13}^2 (1 - r_{12}^2)
\]

Similarly alternative formulas for \( 1.24 \) and \( 1.34 \) can be computed. The following formula can be used to determine a multiple correlation coefficient with three independent variables.

\[
R_{1.24} = \sqrt{(1 - r_{14}^2)(1 - r_{134}^2)(1 - r_{1234}^2)}
\]

Multiple correlation analysis measures the relationship between the given variables. In this analysis, the degree of association is measured between one variable (which is considered as the dependent variable) and a group of other variables (which are considered as independent variables).
Example:

The following are the zero order correlation coefficients

\[ R_{12} = 0.98; \quad r_{13} = 0.44; \quad r_{23} = 0.54. \]

Calculate multiple correlation coefficient treating first variable as dependent and second and third variables as independent.

Solution: First variable is dependent. Second and third variables are independent. Using the formula for multiple correlation coefficients for \( R_{1.23} \) we get.

\[
R_{1.23} = \sqrt{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}/(1-r_{23}^2)} = 0.986
\]

Hence, the multiple correlation coefficient is 0.986.

**Partial correlation coefficient:** Another measure of importance in a multivariate problem is the partial correlation.

**Definition:**

Partial correlation coefficient is a measure of the degree of linear association between any two variables out of a set of variables, when the influence of the remaining variables is eliminated from both of them.

Suppose there are \( K \)-variables \( X_1, X_2, \ldots, X_K \). We want to know the degree of relationship between \( X_1 \) and \( X_2 \), which is free from the influence of \( X_3, X_4, \ldots, X_K \). It is measured by partial correlation coefficient and is denoted by \( \rho_{1234...K} \). Similarly, if we found out the partial coefficient between \( X_2 \) and \( X_4 \), eliminating the influence of the variable \( X_1, X_3, X_5, \ldots, X_K \), it will be denoted by \( \rho_{24135...K} \). The range of partial correlation coefficient is from -1 to +1. Here, we give the formulae for partial correlation coefficient only in case when \( k=2 \) and \( k=4 \).

Suppose, there are three variable \( X_1 \), \( X_2 \) and \( X_3 \). The partial correlation coefficient between \( X_1 \) and \( X_2 \) eliminating the influence of \( X_3 \) is given as

\[
\rho_{123} = \frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{1-\rho_{13}^2}\sqrt{1-\rho_{23}^2}}
\]

Let the estimate of \( \rho_{123} \) be given by \( r_{123} \). To find \( r_{123} \), we replace all simple population correlation coefficients by their estimated values. Thus

\[
r_{123} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}
\]

\( R_{123} \) is also known as the first order partial correlation coefficient.
Polynomial regression model and their fitting

Multiple regression model

Multiple regression analysis is an extension of two variable regression analysis. In this analysis, two or more independent variables are used to estimate the values of a dependent variable, instead of one independent variable.

Objective of multiple regression analysis are:

To derive an equation, which provides estimates of the dependent variable from values of the two or more independent variables.
To obtain the measure of the error involved in using the regression equation as a basis of estimation.
To obtain a measure of the proportion of variance in the dependent variable accounted for or explained by the independent variables.

Multiple regression equation explains the average relationship between the given variables and the relationship is used to estimate the dependent variable. Regression equation refers the equation for estimating a dependent variable. Estimating dependent variable $X_1$ from the independent variables $X_2$, $X_3$,....., is known as regression equation of $X_1$ on $X_2$, $X_3$.....

Regression equation, when three variables are involved, is given below:

$$X_{1.23} = a_{1.23} + b_{1.23}X_2 + b_{13.2}X_3$$

Where $X_{1.23}$ is an estimated value of the dependent variable, $X_2$ and $X_3$ are independent variable.

$a_{1.23} = \text{(constant) the intercept made by the regression plan. It gives the value of the dependent variable, when all the independent variable assumes a value equal to zero.}$

$b_{1.23}$ and $b_{13.2} = \text{partial regression coefficients or net regression coefficients.}$

$b_{1.23} = \text{measures of the amount by which a unit change in } X_2 \text{ is expected to affect } X_1 \text{ when } X_3 \text{ is held constant.}$

Deviation taken from actual means are
\[ X_{1.23} = b_{1.23}X_2 + B_{13.2}X_3 \]
\[ X_1 = (X_1 - \bar{X}_1) \]
\[ X_2 = (X_2 - \bar{X}_2) \]
\[ X_3 = (X_3 - \bar{X}_3) \]

\[ b_{1.23} \text{ and } b_{13.2} \text{ can be obtained by solving the following equations} \]

\[ \Sigma X_1X_2 = b_{1.23}X_2^2 + b_{13.2}X_2X_3 \]
\[ \Sigma X_1X_2 = b_{1.23} \Sigma X_2X_1 + b_{13.2} \Sigma X_3 \]

\[ b = \frac{\sigma_{1.23}}{\sigma_{3.12}} \]

\[ (X_1 - \bar{X}_1) = \left[ \frac{r_{12} - r_{13} r_{23}}{1 - r_{23}^2} \right] \left[ \frac{S_1}{S_2} \right] + (X_2 - \bar{X}_2) \left[ \frac{r_{12} - r_{13} r_{23}}{1 - r_{23}^2} \right] \left[ \frac{S_1}{S_3} \right] \left( X_3 - \bar{X}_3 \right) \]

Regression equation of \( X_3 \) and \( X_2 \) and \( X_1 \) is:

\[ (X_3 - \bar{X}_3) = \left[ \frac{r_{23} - r_{13} r_{12}}{1 - r_{23}^2} \right] \left[ \frac{S_3}{S_2} \right] + (X_2 - \bar{X}_2) \left[ \frac{r_{13} - r_{23} r_{12}}{1 - r_{23}^2} \right] \left[ \frac{S_3}{S_1} \right] \left( X_1 - \bar{X}_1 \right) \]

**Application of Multiple Regression**

Multiple regression analysis can be applied to test the factor such as export elasticity, import electricity and structural change (contribution of manufacturing sector towards GDP) influencing over employment. Here, employment is a dependent variable.

Similarly, researchers can attempt to sue multiple regression in their research work appropriately.

Solved Example:

Find Karl Pearson’s correlation coefficient for the data displayed in Table given below:

<table>
<thead>
<tr>
<th>X</th>
<th>20</th>
<th>16</th>
<th>12</th>
<th>8</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>22</td>
<td>14</td>
<td>4</td>
<td>12</td>
<td>8</td>
</tr>
</tbody>
</table>
Solution: The table given below displays the sums calculated for the data represented above.

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>X^2</th>
<th>Y^2</th>
<th>XY</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>22</td>
<td>400</td>
<td>484</td>
<td>440</td>
</tr>
<tr>
<td>16</td>
<td>14</td>
<td>256</td>
<td>196</td>
<td>224</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>144</td>
<td>16</td>
<td>48</td>
</tr>
<tr>
<td>8</td>
<td>12</td>
<td>64</td>
<td>144</td>
<td>96</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>16</td>
<td>64</td>
<td>32</td>
</tr>
</tbody>
</table>

\[ \sum X = 60 \quad \sum Y = 60 \quad \sum X^2 = 880 \quad \sum Y^2 = 904 \quad \sum XY = 840 \]

Solution: Applying the formula for ‘r’ and substituting the respective values from the table we get \( r \) as:

\[
\begin{align*}
r & = \frac{\sum \Sigma XY - \Sigma X \Sigma Y}{\sqrt{\left(\sum \Sigma X^2 - (\Sigma X)^2\right) \left(\sum \Sigma Y^2 - (\Sigma Y)^2\right)}} \\
& = \frac{5(840) - (60)(60)}{\sqrt{5(840) - (60)^2} \cdot \sqrt{5(904) - (30)^2}} = 0.70
\end{align*}
\]

Hence, Karl Pearson’s correlation coefficient is 0.70.

**Self assessment questions**

i. From the following data, calculate the correlation between variables 1 and 2 keeping the 3rd constant \( r_{12} = 0.7; r_{13} = 0.6, r_{23} = 0.4 \)

ii. Calculate \( r_{23.1} \) and \( r_{13.2} \) from the following:

\[
r_{12} = 0.60; r_{13} = 0.51, r_{23} = 0.4
\]

iii. Give the zero order correlation coefficients, calculate the partial correlation between variables 1 and 3 keeping the 2nd constant. Interpret your result.

\[
r_{12} = 0.80; r_{13} = 0.6, r_{23} = 0.5
\]

**State whether the following questions are true or false.**

i. Scatter diagram does not give us a quantitative measure of correlation coefficient.

ii. Correlation studies estimate the values of one variable from the knowledge of the other.

iii. Correlation coefficient is an absolute measure.

iv. The correlation studied between height and weight, keeping age as constant.

v. Correlation coefficient is a geometric mean between regression coefficients.

vi. The regression lines pass through \((\bar{X}, \bar{Y})\).

vii. \( b_{yx} = r \cdot \text{S.D. of } X / \text{S.D. of } Y \).

viii. The higher the angle between regression coefficients, the lower the correlation.
Unsolved questions

1. Test the significance correlation for the value based on the number of observations
   i. 10  ii. 100 and ‘r’ is 0.4 and 0.9

2. The following table gives marks obtained by 10 students in agriculture and statistics. Calculate the rank correlation.

   Table: Marks of students obtained in agriculture and statistics
   Marks in Statistics: 35 90 70 40 95 45 60 85 80 50
   Marks in agriculture: 45 70 65 30 90 40 50 75 85 60

3. Calculate Spearman’s rank correlation coefficient between the series A and B given in table.

   Series A: 57 59 62 63 64 65 55 58 57
   Series B: 113 117 126 126 130 129 111 116 112

4. For the data given in the following table, obtain the two lines of regression and its estimation of the blood pressure when age is 50 years.

   Age (x) in years: 56 42 72 39 63 47 52 49 40 42 68 60
   BP (Y): 127 112 140 118 129 116 130 125 115 120 135 133

5. The Table given below displays the results that were worked out from scores in statistics and mathematics in a certain examination.

   Scores in Statistics (X)  Scores in Mathematics (Y)
   Mean: 39.5 47.5
   Standard deviation: 10.8 17.8

   Karl Pearson’s correlation coefficient between X and Y = 0.42. Find both the regression lines. Use these lines to estimate the value of Y when X = 50 and the value of X when Y=30
Answer to SAQs and unsolved Questions

1. i., ii., iii. Refer to text material
2. i. True    ii. False   iii. False   iv. True
3. i. True    ii. True    iii. False   iv. True

Answer to terminal questions

1. i. Non significant     ii., iii. iv. Highly significant
2. 0.903
3. 0.967
4. $X = -95 + 1.184, Y = 87.2 + 0.724$
5. $X = 27.62 + 0.25Y, Y = 20.24 + 0.69X$
CHAPTER 11

Nonparametric Tests

The hypothesis about the parameter (s) of the distribution (s) were tested and the decision about the hypothesis taken with the help of the critical region which is determined by the distribution is not known, one needs statistical methods which do not require the form of the parent distribution. Such methods are called nonparametric or distribution-free methods.

The desired significance level ‘\( \alpha \)’ in nonparametric tests is called the nominal \( \alpha \).

Treatment of the ties in Rank tests

The nonparametric tests are based on the ranks of the ordered observations of a sample. Even though the population has been assumed continuous, all observations are not distinct owing to rounding of values.

Assumptions

1. For nonparametric tests, the only assumption is about the continuity of the distribution function. This is postulated to determine the sampling distributions.
2. Median is as good as index of central tendency as mean. We know, for symmetrical distributions, mean and median coincide. Hence, in nonparametric statistics median is taken as a measure of location parameter instead of mean.

One-sample nonparametric tests

A random sample of size \( n \) is drawn from a population and the sample values are arranged in order of magnitude and ranked accordingly, if need be. Different tests evolved in one sample case for the test of hypothesis.

Kolmogorov-Smirnov (K-S) one Sample test

A random sample \( X_1, X_2, \ldots, X_n \) of size \( n \) is drawn from an unknown population having the cumulative distribution function (c.d.f.) \( F(x) \). Let the ordered valued be \( x_{(1)}, x_{(2)}, \ldots, x_{(n)} \). The step function \( S_n(x) \), with jumps occurring at the values of the ordered statistics \( x_{(1)}, x_{(2)}, \ldots, x_{(n)} \) for the sample, approaches the true distribution for all \( x \).

Making use of this theorem, the comparison of the empirical distribution functions \( S_n(x) \) of the sample for any value \( x \) is made with the population c.d.f. under \( H_0 \), \( F_0(x) \). This comparison is made by defining the distance between the two cumulative distribution
functions which is taken as the supermom of the absolute deviations i.e sup | S_n(x) – F_0(x) | over all x. The hypothesis for the test of goodness of fit is

$$H_0 : F(x) = F_0(x) \text{ vs. } H_1 : F(x) \neq F_0(x)$$

Where, F_0 is a completely specified continuous distribution.

To test $H_0$, the actual numerical difference $| S_n(x) – F_0(x) |$ is used in K-S test. Since this difference depends on $x$, the K-S statistics is taken to be the supremum of such difference i.e.

$$D_n = \sup_{x \text{ overall}} | S_n(x) – F_0(x) |$$

Where, D_n is known as the K-S statistic.

Example : On tossing five coins 192 times, the frequencies of 0 to 5 heads are:

<table>
<thead>
<tr>
<th>No. of heads</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>6</td>
<td>26</td>
<td>73</td>
<td>66</td>
<td>14</td>
<td>7</td>
</tr>
</tbody>
</table>

The fairness of the coin by the K-S test can be ascertained. The null hypothesis under test is,

$$H_0 : F(x) = F_0(x) \text{ vs. } H_1 : F(x) \neq F_0(x)$$

The hypothetical frequencies are calculated with the help of the binomial function. Here, $\binom{n}{r} p^r q^{n-r}$ n = 5, p = 0.5.

Thus, the frequencies for r = 0, 1, 2, 3, 4, 5 are

$$f_0 = \binom{5}{0} \left(\frac{1}{2}\right)^5 \times 192 = 6, \quad f_1 = \binom{5}{1} \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^1 \times 192 = 30,$$

and similarly, $f_2 = 60, f_3 = 60, f_4 = 30, f_5 = 6$. Actual and theoretical frequencies, the sample or empirical c.d.f., theoretical c.d.f. and the differences are tabulated below.

<table>
<thead>
<tr>
<th>No. of heads</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual frequencies</td>
<td>6</td>
<td>26</td>
<td>73</td>
<td>66</td>
<td>14</td>
<td>7</td>
</tr>
<tr>
<td>Empirical c.d.f.</td>
<td>6/192</td>
<td>32/192</td>
<td>105/192</td>
<td>171/192</td>
<td>185/192</td>
<td>192/192</td>
</tr>
<tr>
<td>Theoretical frequencies</td>
<td>6</td>
<td>30</td>
<td>60</td>
<td>60</td>
<td>30</td>
<td>6</td>
</tr>
</tbody>
</table>
Theoretical c.d.f.  
\begin{align*}
6 & 32 & 105 & 171 & 185 & 192 \\
192 & 192 & 192 & 192 & 192 & 192 
\end{align*}

Difference : \( D_n \)  
\begin{align*}
0 & 4 & 9 & 15 & 1 & 0 \\
192 & 192 & 192 & 192 & 192 & 192 
\end{align*}

\[ \text{Max } D_n = \frac{15}{192} = 0.078 \]

Suppose the prefixed significance level \( \alpha = 0.1 \). The critical value of \( D_n \) from Table is \( \frac{1.22}{\sqrt{192}} = 0.088 \). Since the value of statistic \( D_n \) does not exceed the critical value 0.088, \( H_0 \) is not rejected. This means that the sample c.d.f. is similar to the hypothetical distribution function.

**The Ordinary Sign Test**

Let a random sample \( x_1, x_2, \ldots, x_n \) of size \( n \) be drawn from a population \( F(x) \) where \( F(x) \) is assumed to be continuous in the close vicinity of the median. Suppose, the median of \( F(x) \) is \( M \). In this situation, \( P(X = M) = 0 \).

The hypothesis which has to be tested by sign test is,

\[ H_0 : M = M_0 \] vs. \[ H_1 : M \neq M_0 \]

Where, \( M_0 \) is a given value of the population median, we know,

\[ P(X > M_0) = P(X < M_0) = 0.5 \]

Hence, the null hypothesis \( H_0 \) under test is equivalent to

\[ H_0 : P(X > M_0) = P(X < M_0) \] vs. \[ H_1 : P(X > M_0) \neq P(X < M_0) \]

If the randomly selected sample comes from a population having the median value equal to \( M_0 \) on the average, half of the observations will be below \( M_0 \) and half above \( M_0 \). To perform the sign test, we take the differences \( (X_i - M_0) \) for \( i = 1, 2, \ldots, n \) and consider their signs. Let the number of positive signs be \( r \) and negative signs \( (n-r) \). For the test statistic, we consider only the positive signs. In this way the data have been dichotomised which consist of the number of positive and negative signs. The distribution of \( r \) given \( n \) is a binomial distribution with \( p = P(X > M_0) \). Thus, the null hypothesis \( H_0 \) changes to \( p = 0.5 \). So now we test,

\[ H_0 : p = 0.5 \] vs. \( H_1 : p \neq 0.5 \]

Alternate hypothesis \( H_1 : p < 0.5 \) will be chosen when it is expected that the sample will have few positive signs.
Wilcoxon Singed Rank Test:

The ordinary sign test was based on only the sign of the deviations of the ordered sample values from the hypothesized median $M_0$. No heed was paid to the magnitude of the differences. The Wilcoxon signed rank test utilizes the signs as well as the magnitudes of the differences. This test is more sensitive and powerful than the ordinary sign test.

The variable $T^+$ is distributed with mean

$$\sum_{i=1}^{n} i \cdot p_i \quad \text{and variance} \quad \sum_{i=1}^{n} t^2_i \cdot p_i \cdot (1 - p_i).$$

Under $H_0$, $p_i = \frac{1}{2}$ and hence,

$$E(T^+) = \frac{1}{2} \sum_{i=1}^{n} p_i - \frac{1}{2} \sum_{i=1}^{n} 1 = \frac{1}{2} \cdot \frac{n(n+1)}{2} = \frac{n(n+1)}{4}$$

$$V(T^+) = \frac{1}{4} \sum_{i=1}^{n} t^2 = \frac{n(n+1)(2n+1)}{24}$$

Run Test

In statistical theory, it has normally be assumed that a sample drawn from a population is random. Whether the assumption of randomness is true or not needs verification. The run test is one device to test randomness. Before discussing the test, it is essential to discuss the run.

Definition: A run is a sequence of like symbols which are followed and preceded by other kinds of symbols or no symbol at either side. For clarity, a vertical line is drawn in between two consecutive sequences of symbols to mark the runs. A sequence of symbols exhibiting a pattern of symbols is usually indicative of lack of randomness.

Test of Randomness:

The null hypothesis, $H_0$: the symbols ‘a’ and ‘b’ occur in random order in the sequence against the alternative, $H_1$: symbols ‘a’ and ‘b’ do not occur in random order; can be tested by the run test. Let the sample of size $n_1$ contains a symbols of one type, say a, and $n_2$ symbols of the other type say b. Thus, $n = n_1 + n_2$. Also suppose the number of runs of symbol ‘a’ and $r_1$ and that of symbol b and $r_2$. Suppose $r_1 + r_2 = r$. In order to perform a test of hypothesis based on the random variable R, we need to know the probability distribution of R under $H_0$. The probability distribution function of R is given as.
\[ f_r(r) = 2 \left( \frac{n_1 - 1}{r - 1} \right) \left( \frac{n_2 - 1}{r - 1} \right) \left( \frac{n_1 + n_2}{n_1} \right) \]

When ‘r’ is even.

For \( r \) even, the number of runs of both types must be the same i.e. \( r_1 = r_2 = r/2 \).

\[ f_r(r) = \left\{ \left( \frac{n_1 - 1}{r - 1} \right) \left( \frac{n_2 - 1}{r - 1} \right) \right\} \left( \frac{n_1 + n_2}{n_1} \right) \]

When ‘r’ is odd.

For \( r \) odd, \( r_1 = r_2 = r \pm 1 \). In this situation, the sum is taken over two pairs of values, \( r_1 = (r-1)/2 \) and \( r_2 = (r+1)/2 \) and vice versa.

Example: Given the following sequence as given in the above text,

\begin{align*}
\text{aa} & \mid \text{bbb} \mid \text{a} \mid \text{bb} \mid \text{aaaa} \mid \text{bb} \mid \text{a} \\
\end{align*}

The null hypothesis \( H_0 \): the observations occur in random order,

Against \( H_1 \): the observations do not occur in random order, can be tested in this manner.

\begin{align*}
\text{n}_1 &= 8, \text{n}_2 = 7, \text{n} = 15 \\
\text{r}_1 &= 4, \text{r}_2 = 3, \text{r} = 7 \\
\text{For } \alpha = 0.05, \text{n}_1 = 8, \text{n}_2 = 7, \text{the lower critical value} = 4 \\
\text{For } \alpha = 0.05, \text{n}_1 = 8, \text{n}_2 = 7, \text{the upper critical value} = 13 \\
\text{The number of runs in the sample is 7 which lies between 4 and 13. It means that} \\
\text{the observations occur in random order with 5 per cent significance level.} \\
\end{align*}

Two sample non parametric test

Median text:

This test is attributed to Westenberg (1948) and Mood (1950). The sign test for two sample problems has been applicable only when the observations are paired. But such a situation is not always possible. Often we have samples of different sizes from two populations. In such a situation the median test is quite-efficient. Suppose two random samples \( X_1, X_2, ..., X_n \) and \( Y_1, Y_2, ..., Y_n \) of size \( n_1 \) and \( n_2 \) are drawn from two populations \( F_x \) and \( F_r \), respectively. Suppose \( n_1 + n_2 = n \). The median test is a test of
equality of location parameters of the two populations under consideration. Here, we test the hypothesis.

\[ F_x(x) = F_y(x) \text{ for all } x, \]

against the shift alternative.

\[ F_x(x) = F_y(x - \delta) \text{ for all } x \text{ and } \delta \neq 0. \]

**Wald-Walfowitz Runs Test:**

This test is used for testing the identity of two populations. To avoid any confusion between one-sample run test and this test procedure, we will consider two random samples of size \( m \) and \( n \), respectively, instead of size \( n_1 \) and \( n_2 \). Let the two independent samples \( X_1, X_2, ..., X_m \) and \( Y_1, Y_2, ..., Y_n \) combine into a single sequence of ordered statistics.

The mean and variance of \( R \) are given as:

\[
\text{mean } (R) = \frac{2mn}{m+n} + 1 \\
\text{and } \text{var } (R) = \frac{2mn(2mn-n-m)}{(m+n)^2 (m+n-1)}
\]

Thus, the normal deviate is,

\[ Z = \frac{R - \text{mean } (R)}{\sqrt{\text{var } (R)}} \]

Where \( Z \sim N (0, 1) \)

The decision about \( H_0 \) is taken in the usual way.

**Example:**

The following are the rates of flow of a certain gas through two soil samples collected from two different places.

Sample X : 23, 27, 19, 24, 22, 30

Sample Y : 21, 29, 34, 32, 26, 28, 36, 26

The hypothesis that the populations of soil types are the same with respect to the rates of flow through the soils, i.e. symbolically,

\[ H_0 : F_x(x) = F_y(x) \text{ for all } x \]
H₁ : Fₓ(x) ≠ Fᵧ(x) for some x

Can be tested by the Wald-Wolfowitz test

The sequence of combined samples with the clearly marked runs is,

19 | 21 | 22, 23, 24 | 26, 26 | 27 | 28, 29 | 30 | 32, 34, 36|

In this case m = 6, n = 8 and the number of runs r = 8. The probability for r = 8 (even) is,

\[ f(r) = \frac{2 \binom{5}{3} \binom{7}{3}}{\binom{14}{6}} \]

\[ = 0.233. \]

Supposing the predecided level of significance = 0.1. Since the probability \( f(r) \) is greater than 0.1, we accept \( H₀ \) which means that the two soils have identical distributions in respect of rates of flow of a certain gas.

**Mann-Whitney U Test**

Just like the Wald-Wolfowitz runs test, the Mann-Whitney U-test is based on the idea that the pattern of X’s and Y’s exhibited by combined ordered statistics provides information about the relationship between their parents populations. Unlike previous runs test, the Mann-Whitney test is based on the criterion of magnitudes of Y’s relation to X’s, or vice versa. If most Y’s are greater or less than X’s, the null hypothesis of equality of two populations is most likely to be refuted.

In the ordered sequences of combined samples, the ranks of either X’s or Y’s are found out. In case, the ranks of Y’s are taken into consideration (the case when Y precedes X) and their sum is \( S₂ \), the Mann-Whitney U-statistic is given as,

\[ U = n₁ n₂ + \frac{n₁(n₁+1)}{2} - S₂ \]

Similarly, if the ranks of X’s are counted i.e. the case when X precedes Y and their sum is \( S₁ \), the statistic U is give as,

\[ U' = n₁ n₂ + \frac{n₂(n₂+1)}{2} - S₁ \]

It is interesting to point out that both the approaches yield the same value of 1/1 or instance consider the ordered sequence given earlier for \( n₁=5 \) and \( n₂=3 \).
Sequence: \( Y_1 \ X_5 \ X_1 \ Y_3 \ X_2 \ X_3 \ X_4 \ Y_2 \)

Ranks: \( 1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \)

First we consider \( D_y = 1 \) where \( Y_j < X_i \), otherwise zero. In the given sequence

- \( Y_1 \) is less than \( X_5, X_1, X_2, X_3, X_4 \), hence \( D_{11} = 5 \)
- \( Y_3 \) is less than \( X_2, X_3, X_4 \), hence \( D_{i3} = 3 \)
- \( Y_2 \) is less than none, hence \( D_{i2} = 0 \)

\[
U = \sum_{j=1}^{3} \sum_{i=1}^{5} D_{ij} = 5 + 3 + 0 = 8
\]

Now, statistic \( U \) can be calculated as follows:

\[
n_1 = 5, \ n_2 = 3, \ S_2 = 1 + 4 + 8 = 13.
\]

\[
U = 5 \times 3 + \frac{3 \times 4}{2} = 13 = 8
\]

Similarly, if we consider \( D_{ij} = 1 \) for \( X_i < Y_j \), other wise zero, we get,

\[
D_{5j} = 1+1 = 2, \ D_{1j} = 1 + 1 = 2, \ D_{2j} = 1, \ D_{3j} = 1, \ D_{4j} = 1
\]

and

\[
U = \sum_{i=1}^{5} \sum_{j=1}^{3} D_{ij} = 2 + 2 + 1 + 1 = 7
\]

\[
U' = 5 \times 3 + \frac{5 \times 6}{2} = 23 = 7
\]

since,

\[
S_1 = 2 + 3 + 5 + 6 + 7 = 23
\]

**Kruskal-Wallis Test:**

This is a nonparametric test for one-way classification but like one-way analysis of variance. The Kruskal-Wallis test is an improvement of the median test. In the median test, the magnitude of the observations was compared with one value only, i.e. the median. But in this method the magnitude of observations is compared with every other observation by considering the ranks.

The Kruskal-Wallis test statistic is

\[
H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N + 1)
\]
The statistic $H$ is distributed as chi-square with $(k-1)$ d.f. Reject $H_0$ at significant level $\alpha$ if,

$$H \geq \chi^2_{\alpha, k-1}$$

Otherwise $H_0$ is not rejected.

Example:
The green pod yield (kg) under four treatments is as tabulated below:

<table>
<thead>
<tr>
<th>No of plots</th>
<th>Treatments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>3.17</td>
</tr>
<tr>
<td>2</td>
<td>3.40</td>
</tr>
<tr>
<td>3</td>
<td>3.50</td>
</tr>
<tr>
<td>4</td>
<td>2.87</td>
</tr>
<tr>
<td>5</td>
<td>3.88</td>
</tr>
<tr>
<td>6</td>
<td>4.00</td>
</tr>
<tr>
<td>7</td>
<td>3.60</td>
</tr>
</tbody>
</table>

The hypothesis that there is no difference among four treatments can be tested by the median test.

Here $k = 4$, $n_1 = 7$, $n_2 = 7$, $n_3 = 5$ $n_4=6$ and $N = 25$. The ordered statistics of the pooled samples is,

$(2.37), (2.48), (2.48), (2.58), 2.69, 2.80, (2.84), 2.87, 2.88, s.97, (3.00), 3.10, 3.15, 3.17, 3.25, 3.27, 3.40, 3.44, 3.45, 3.50, 3.60, 3.87, 3.88, 0.94, 4.00.$

$M_d = 3.15$

For treatment 1; $U_1 = 1$

For treatment 2; $U_2 = 2$

For treatment 3; $U_3 = 3$

For treatment 4; $U_4 = 6$
\[
r = \sum_{i=1}^{4} U_i = 1 + 2 + 3 + 6 = 12
\]

To show the method of calculations, the test will be performed at 5 per cent significance level by using (i) hyper-geometric distribution and (ii) chi-square statistic

i. The probability

\[
(U_1, U_2, U_3, U_4 / r) = \left( \begin{array}{cccc}
\frac{7}{1} & \frac{7}{2} & \frac{5}{3} & \frac{6}{6} \\
\frac{25}{12} & & & \\
& & & \\
& & & \\
& & & \\
& & & \\
& & & \\
& & & \\
& & & \\
& & & \\
\end{array} \right)
\]

\[
= \frac{21}{74290} = 0.00028
\]

The calculated probability is less than 0.05. Hence, we reject H0 which means that the treatments differ significantly.

(ii) Observed frequencies in two categories are:

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Category I (Less than ( \theta ))</th>
<th>Category I (More than ( \theta ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 ( \frac{12}{25} \times 7 = 3.4 )</td>
<td>6 ( \frac{13}{25} \times 7 = 3.6 )</td>
</tr>
<tr>
<td>2</td>
<td>2 ( \frac{12}{25} \times 7 = 3.4 )</td>
<td>5 ( \frac{13}{25} \times 7 = 3.6 )</td>
</tr>
<tr>
<td>3</td>
<td>3 ( \frac{12}{25} \times 5 = 2.4 )</td>
<td>1 ( \frac{13}{25} \times 5 = 2.6 )</td>
</tr>
<tr>
<td>4</td>
<td>6 ( \frac{12}{25} \times 6 = 2.8 )</td>
<td>0 ( \frac{13}{25} \times 6 = 3.12 )</td>
</tr>
</tbody>
</table>

By formula (11.28) \( x^2 = 11.85 \)

By formula (11.28.1) = \( \frac{25 \times 25}{12 \times 13} (0.796 + 0.264 + 0.072 + 1.622) = 11.03 \)

Tabulated value of \( \chi^2_{0.05,3} = 7.815 \)

Again, the conclusion is the same as in procedure
Friedman’s Test:

This is a nonparametric test of k-related samples of equal size say n, parallel to two-way analysis of variance. Here we have k-related samples of size n arranged in n blocks and k columns in a two-way table as given below:

<table>
<thead>
<tr>
<th>Blocks</th>
<th>Samples (Treatments)</th>
<th>Block totals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>R11 R12...R1k</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>R21 R32...R2k</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>R31 R32...R3k</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rn1 Rn2...Rnk</td>
<td></td>
</tr>
<tr>
<td>n</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Col. Totals</td>
<td>R1 R2...Rk</td>
<td>(\frac{nk(k+1)}{2})</td>
</tr>
</tbody>
</table>

Where, \(R_{ij}\) is the rank of the observation belonging to sample \(j\) in block \(i\) for \(j = 1, 2, ..., k\) and \(i = 1, 2, ..., n\). This is the same situation in which there are \(k\) treatments and each treatment is replicated \(n\) times. Here it should be carefully noted that the observations in a block receive ranks from 1 to \(k\). The smallest observation receives rank 1 at its place and the largest observation receives rank \(k\) at its place. Intermediary observations receive ranks accordingly. Hence, the block totals are constant and equal to \(\frac{nk(k+1)}{2}\), the sum of \(k\) integers.

The null hypothesis \(H_0\) to be tested is that all the \(k\) samples have come from identical populations. In case of experimental design, the null hypothesis \(H_0\) is that there is no difference between \(k\) treatments. The alternative hypothesis \(H_1\) is that at least two samples (treatments) differ from each other.

Under \(H_0\), the test statistics is

\[
F = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1)
\]
The statistics $F$ is distributed as chi-square with $(k-1)$ d.f. At level ‘$\alpha$’, reject $H_0$ if $F \geq \chi^2_{\alpha,k-1}$. Otherwise $H_0$ is not rejected.

**Example:**

The iron determinations (ppm) in five pea-leaf samples, each under three treatments, were as tabulated below:

<table>
<thead>
<tr>
<th>Blocks</th>
<th>Treatments</th>
<th>Block total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>591 682 727</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>(1) (2) (3)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>818 591 863</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>(2) (1) (3)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>682 636 773</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>(2) (1) (3)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>499 625 909</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>(1) (2) (3)</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>648 863 818</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>(1) (3) (2)</td>
<td></td>
</tr>
<tr>
<td>Col. Totals</td>
<td>1 9 14 30</td>
<td></td>
</tr>
</tbody>
</table>

The null hypothesis, $H_0$, that the iron content in leaves under three treatments in the same against $H_1$, that at least two of them have different effect, can be tested by Friedman’s test. The ranks of observations in each block are shown below them in parentheses. Friedman’s test statistics by formula is,

$$F = \frac{12}{5 \times 3 \times 4} \left( \frac{7^2 + 9^2 + 14^2}{3 \times 5 \times 4} \right)$$

For $\alpha = 0.05$, the table values of $\chi^2_{0.05,2} = 5.99$. Since the calculated $F$-value is less than 5.99, $H_0$ is not rejected. This means that there is no difference in the iron content of pea leaves due to the treatments.
Kendall's $W$

It is also known as Kendall's coefficient of concordance, a non-parametric statistic. It is a normalization of the statistic of the Friedman test, and can be used for assessing agreement among raters. Kendall's $W$ ranges from 0 (no agreement) to 1 (complete agreement).

Suppose, for instance, that a number of people have been asked to rank a list of political concerns, from most important to least important. Kendall's $W$ can be calculated from these data. If the test statistic $W$ is 1, then all the survey respondents have been unanimous, and each respondent has assigned the same order to the list of concerns. If $W$ is 0, then there is no overall trend of agreement among the respondents, and their responses may be regarded as essentially random. Intermediate values of $W$ indicate a greater or lesser degree of unanimity among the various responses.

While tests using the standard Pearson correlation coefficient assume normally distributed values and compare two sequences of outcomes at a time, Kendall's $W$ makes no assumptions regarding the nature of the probability distribution and can handle any number of distinct outcomes.

$W$ is linearly related to the mean value of the Spearman's rank correlation coefficients between all pairs of the rankings over which it is calculated.

**Definition**

Suppose that object $i$ is given the rank $r_{i,j}$ by judge number $j$, where there are in total $n$ objects and $m$ judges. Then the total rank given to object $i$ is

$$R_i = \sum_{j=1}^{m} r_{i,j},$$

and the mean value of these total ranks is

$$\bar{R} = \frac{1}{2}m(n + 1).$$

The sum of squared deviations, $S$, is defined as

$$S = \sum_{i=1}^{n}(R_i - \bar{R})^2,$$

and then Kendall's $W$ is defined as$^{[1]}$
If the test statistic $W$ is 1, then all the judges or survey respondents have been unanimous, and each judge or respondent has assigned the same order to the list of objects or concerns. If $W$ is 0, then there is no overall trend of agreement among the respondents, and their responses may be regarded as essentially random. Intermediate values of $W$ indicate a greater or lesser degree of unanimity among the various judges or respondents.

Legendre discusses a variant of the $W$ statistic which accommodates ties in the rankings and also describes methods of making significance tests based on $W$.

**Correction for ties**

When tied values occur, they are each given the average of the ranks that would have been given had no ties occurred. For example, the data set $\{80,76,34,80,73,80\}$ has values of 80 tied for 4th, 5th, and 6th place; since the mean of $\{4,5,6\} = 5$, ranks would be assigned to the raw data values as follows: $\{5,3,1,5,2,5\}$.

The effect of ties is to reduce the value of $W$; however, this effect is small unless there are a large number of ties. To correct for ties, assign ranks to tied values as above and compute the correction factors

$$T_j = \sum_{i=1}^{g_j}(t_i^3 - t_i),$$

where $t_i$ is the number of tied ranks in the $i$th group of tied ranks, (where a group is a set of values having constant (tied) rank,) and $g_j$ is the number of groups of ties in the set of ranks (ranging from 1 to $n$) for judge $j$. Thus, $T_j$ is the correction factor required for the set of ranks for judge $j$, i.e. the $j$th set of ranks. Note that if there are no tied ranks for judge $j$, $T_j$ equals 0.

With the correction for ties, the formula for $W$ becomes

$$W = \frac{12\sum_{i=1}^{n}(R_i^2) - 3m^2n(n+1)^2}{m^2n(n^2-1) - m\sum_{j=1}^{m}(T_j)} ,$$

where $R_i$ is the sum of the ranks for object $i$, and $\sum_{j=1}^{m}(T_j)$ is the sum of the values of $T_j$ over all $m$ sets of ranks.
CHAPTER 12

Probit Regression Analysis

A great deal of biological work is now concerned with the assay of drugs, vitamins, sera and other forms of stimuli. In these experiments the effect of stimulus is determined according to the reaction of living organisms. Usually the stimulus is applied at a series of levels and the reaction of each level determined from its application to a batch of subjects.

Instead of speaking in general terms it is more convenient, in describing the kind of data obtained, to refer to one type of experiment such as the determination of the toxicity of a chemical preparation to a given type of insect. In such an experiment various concentrations are prepared and a batch of insects assigned at random to each concentration level. The substance is applied, and for each batch a count is made of the total number of insect (n) and the number killed (r). The results can be expressed either as a proportion (r/n) or as a percentage 100(r/n). This is the type of data to which the methods of probit analysis are appropriate in assessing the value of different toxic substances, comparing relative potencies of different substances, applying tests of significance, and determining fiducial limits.

The distribution of tolerance:

For any one subject there is a level of intensity of the stimulus below which response does not occur and above which it does occur. This level is referred to as the tolerance for that subject. For example, in considering a given insect there is a level of concentration for a certain poison such that if the concentration is less than this level the insect will live, but if the concentration is greater the insect will die. Let this tolerance of a subject be represented by \( \lambda \); then in a population of subjects we are concerned with the distribution of \( \lambda \). It is generally true for most biological preparations that the distribution of \( \lambda \) is not
normal but is at least approximately normal for \( X = \log_{10} \lambda \). In the usual analysis, therefore, we deal with the log of the concentrations. \( \lambda \) is referred to as the dose in terms of actual concentration such as milligrams per liter, and the \( \log^{10} \) concentration is referred to as the dosage or dose metameter.

It should be noted that the log transformation does not necessarily normalize the distribution of tolerances. In certain cases other transformations are more appropriate, but the log transformation is found to apply in most experiments and there are chemical and biological reasons why this should be true.

**The probit transformation:**

Assuring that the dosage in a population is normally distributed, we can picture the situation as in Fig. - 1. In the upper portion of the figure we have a normal curve in which the dosage deviations from the mean \((X - \bar{X})\) are replaced on the base line by \((X - \bar{X})/\sigma\) or normal equivalent deviate (N.E.D.). In other words this can be regarded as normal distribution with unit standard deviation. Above the graph of the normal distribution are the percentages of the total area from a given point on the base line to N.E.D. = - \( \infty \). For example, at N.E.D. = -1, the area to the left of that point is 15.87% of the total. These percentages are represented in the lower half of the figure by the sigmoid curve.

It should be clear from the figure that, if we have a population in which the tolerance for a toxic substance is normally distributed and we apply the substance at a series of levels to different batches, the percentages killed when plotted against dosage should give a sigmoid curve. The analysis of results from the sigmoid curve present some rather serious difficulties; therefore, it would seem to be desirable to use a transformation of the percentages such that with a normal distribution the transformed percentages would lie on a straight line. The obvious transformation for this purpose is the corresponding N.E.D. Thus, on obtaining a percentage kill at a given level of 15.87% we would locate in a table the N.E.D. such that if the distribution is normal a kill of exactly
15.87% would be expected. The value would be \(-1\). For a kill of 2.27% the N.E.D. would be \(-2\) and so forth.

The suggestion that the N.E.D. be used as a transformation for percentage response seems to have been first made by Fechner in 1860, but is was not considered seriously until it was again made by Gaddum in 1933. Later in 1934, Bliss suggested adding 5 to the N.E.D. in order to remove negative numbers, and he also suggested that term probit. At that time and since then Dr. Bliss has been responsible for a great deal of the development work with the probit transformation. At the present the best summary of the theory and application of probit analysis in *Probit Analysis* by D. J. Finney.

In Figure-1 the probit scale is shown on the right of the lower part of the figure and the straight line represents the probits of the percentages at all point along the normal curve.

![Diagram](image)

**Fig- 1 :** Theoretical distribution of tolerance and relation of percentage kill to probits
It should be now be clear that, in a typical experiment where in the tolerance are normally distributed, a graph of the probits corresponding to determined percentages will tend to be straight line. Any variation from the normal curve will cause the plotted probits to vary from a straight line. Generally, the observed variations from a straight line are of two types. In the first place the batches of subjects may not be all uniform or the conditions for the batches may not be uniform. This will tend to produce an abnormal scatter of the points about the straight line. In the second place the transformation of the dose to dosage may not be suitable. Usually this will be indicated by trends towards curvi-linear rather than linear regression.

Practical applications of probit analysis:

Since we expect to get a straight-line graph when probits are plotted against dosage, the methods of linear regression are suggested. What we have to decide however, is whether or not such methods are applicable: if they will yield estimates of required population parameters, if valid tests of significance can be applied, and if fiducial limits can be determined.

First we require a measure of the potency of the preparation. It has been concluded generally that the dose giving a 50% kill is the most valuable statistic. It is referred to as the LD 50, (median lethal dose). In experiments here the response is not death we refer to the ED 50 (median effective dose). Whatever practical advantages there may be in knowing the LD 90 or some similar value, the fact is that much greater precision can be obtained in the measurement of the LD 50. This is obvious when we consider the measurement of a very high dosage such as LD 100. Any levels administered beyond this point would give no information whatever. In the case of LD 50, levels above and below contribute equally to the results.

Another factor to be measured is the range of the dosage required for a given range of percentage kill. This might be referred to as the sensitivity of the preparation tested. Obviously if small changes in concentrations give a wide range in the percentage kill, the sensitivity is high.
Referring now to a probit graph such as Figure -1, which will be described in detail later, it is clear that the LD 50 with respect to dosage is given by the dosage that corresponds to a probit value of 5.0. Therefore, if a straight line can be fitted, the LD 50 can be read directly from the graph.

Again on referring to the graph it is evident that sensitivity will be represented by the slope of the line. The greater the slope, the narrower the range in dosage for a given range in the percentage kill.

The geometry of the line would seem to give us therefore the required measures of potency and sensitivity. Locating 2 points \( X_1 \) and \( X_2 \), representing dosage, on the abscissa of the graph and finding the corresponding points \( Y_1 \) and \( Y_2 \) on the probit scale will give the slope of the line. If \( b \) represents the slope we have

\[
b = \frac{Y_2 - Y_1}{X_2 - X_1}
\]

This makes it possible to set up a regression equation of the type \( Y = a + bX \), where \( a = Y_1 - bX_1 \) or \( Y_2 - bX_2 \). This line can also be used to locate the LD 50 which we now represent by the symbol \( m \). Putting \( Y = 5 \) and \( m \) for \( X \), the regression equation is \( 5 = a + bm \), which can be solved by \( m \).

**Fitting a probit regression line:**

The fitting of suitable straight-line regression would seem from the argument above to provide estimates of the required parameters. On making a close study of the problem of fitting such a line, however, we find that there is an essential difference between fitting a line to probit values and fitting a regression line to measure the relation between 2 variables in the ordinary case. Actually, the variance is a minimum at LD 50 and goes to infinity, on one end at the 100% kill, and on the other end at the 0% kill. In order to fit a regression line accurately it is necessary to weight the values at each point by the inverse
of the variance. It has been shown that if $P$ represents the probability of kill at a given dosage level and $Q = (1 - P)$, the probability of survival, the correct weighting coefficient $w = Z^2/PQ$, where $Z$ is the ordinate of the normal distribution corresponding to the probability $P$. Taking $Y$ as a probit value on the straight line, the values of $Z^2/PQ$ have been tabulated by Bliss and reproduced by Fisher and Yates. To illustrate the use of the weighting coefficients we must calculate the mean $X$ from the formula $\Sigma(nwX)/\Sigma(nw)$ in place of usual formula $\Sigma(nX)/\Sigma(n)$, where $n$ represents the total number of insects in a batch.

One difficulty that arises in applying the weighting coefficient is that they are based on $Z$, $P$ and $Q$, which are parameters of the population and not estimates for the data of the experiment. In other words we must theoretically know the equation for the straight line in order to determine the values of $w$, which of course is impossibility. In actual practice two things can be done to overcome the difficulty. First, many examples are such that a sufficiently good straight line can be fitted by obtaining what is known as an eye fit. A straight edge is laid along the graph of probit points, and, if it is obvious that a close fit to the points can be obtained, this line is drawn and used for obtaining the required measures of potency and sensitivity, and further refinement is unnecessary. Second, a line fitted by eye (provisional line) will provide first approximations of $Y$ and consequently of $w$. Substituting the approximate values of $Y$ and $w$, another line is fitted which is a better fit than the provisional line. If further refinement is required, a third line can be fitted using the second line to provide estimates of $Y$ and $w$. This process, if repeated, will lead very quickly to a line that does not change on further applications of the fitting process. This is known as an iterative process, and by suitable methods of fitting it can be made to give the maximum likelihood solution. It may appear to be tedious, but actual practice there is little need to go beyond the second line.

The details of fitting and the formulas required are best observed from the actual examples given below. The same is true for learning the meaning of the terms and the interpretation of the data.
Example 1. Fitting a provisional probit regression line. Table -1 is taken from data given by Morrison for the effect of different concentrations of nicotine sulphate in a 1% saponin solution, on *Drosophila melanogaster*. The various steps in the calculations are best carried out as enumerated below:

1. Set up the data as in the first 3 columns of Table -1 and calculate % kill as in column (4).
2. Determine log concentrations after multiplying the concentrations by a suitable factor to remove negative values. These are entered in column (5).
3. Look up empirical probits in tables by Fisher and Yates, entering the table under % kill. Record these in column (9).
4. Draw Figure -2. First represent the empirical probits by dots and then drawn the straight line to give a good eye-fit. In this example it is fairly easy to draw a suitable line. One point that should be kept in mind when it is difficult to decide where the line should be placed is that the most attention should be paid to those points representing kills of 40 to 60%, and percentage kills outside of the limit of 16 to 84 should be practically disregarded.
5. From the regression line determine:
a. The LD 50 and LD90. The LD 50 corresponds to a probit of 5.0, and the LD 90 to a probit of 6.28. In this figure we get \( m = LD \ 50 = 137 \) an LD 90 = 1.86.

b. The regression coefficient. Locate 2 convenient points at the ends of the line and write down corresponding values of X and Y. Thus

\[
X_1 = 1.0 \quad Y_1 = 4.02 \\
X_2 = 2.0 \quad Y_2 = 6.66
\]

Then the increase in Y for unit increase in X is

\[
X = \frac{Y_2 - Y_1}{X_2 - X_1} = 2.64 = b,
\]

6. In the regression equation obtain values of Y falling on the straight line for the levels of X require. The X’s are entered in column (1) of Table -2, and the Y’s column (2). If the figure is accurately drawn, the Y values can be read from the figure instead of calculating them from the equation.

7. Reading backwards from Fisher and Yate’s table of probits, find the corresponding values of P which are entered in column (3). Then complete the entries in columns (4) and (5), copying from Table -1.

8. Calculate the entries in column (6), (7) and (8) indicated. Note the P is a proportion and not a percentage. Thus a percentage of 16.4 is entered as 0.164. Add the last column to obtain \( \chi^2 = 0.769 \). This \( \chi^2 \) is based on 3 degrees of freedom since we have adjusted for m and b. From Table we note that a \( \chi^2 \) of 0.769 for 3 degree of freedom corresponds to a probability of about 0.85. This is a better fit than can ordinarily be expected but is not sufficiently close to lead us to be suspicious of the data.

9. Test of significance can now be applied to b and m, and fiducial limits calculated. The required preliminary calculations are shown in Table -3. Note that Y is entered to the first decimal place only. The value of Y can be read to the second decimal place from the regression graph, but it can be shown that such a degree of accuracy is unnecessary at this point. The fourth column contains the weighting
coefficients that are determined from Fisher an Yates Table XI. We require the sum of squares of X given by

\[ \sum (nwx^2) = \sum (nwX^2) - \frac{\sum (nwX^2)}{\sum (nw)} = 31.5319 \]

Then the variance of b is

\[ V_b = \frac{1}{\sum (nwx^2)} = \frac{1}{31.5319} = 0.031714 \]

then the fiducial limits are given by \( b \pm t_s b \)

where, \( t = 1.96 \). If \( \chi^2 \) had indicated significant heterogeneity, it would have been necessary to make a correction in obtaining the fiducial limits. This will be demonstrated after first making the usual calculation of the limits.

\[ 2.64 \pm 1.96 \times 0.178 = 2.64 \pm 0.35 \]

giving 2.29 to 2.99 as the required limits.

If heterogeneity is present, we calculate the heterogeneity factor

\[ \mu = \frac{\chi^2}{DF} \]

where the degrees of freedom = \( k - 2 \), \( k \) being the number of batches tested. Then the variance of b is

\[ V'b = \frac{\mu}{\sum (nwx^2)} \]
and

\[ S'_b = \sqrt{V'_b} \]

The fiducial limits are then

\[ b \pm t s'_b \]

Where \( t \) is the value required for significance at the 5% level for \( k-2 \) degrees of freedom.

To get the fiducial limits of \( m \) we calculate

\[
V_m = \frac{1}{b^2} \left[ \frac{1}{\sum(nw)} + \frac{(m - \bar{X})^2}{\sum(nwx^2)} \right]
\]

\[
= \frac{1}{2.64^2} \left[ \frac{1}{310.46} + \frac{(1.37 - 1.56)^2}{31.5319} \right]
\]

\[
= 0.000 \, 626 \, 4
\]

\[ S_m = \sqrt{0.000 \, 626 \, 4} = 0.025 \]

The fiducial limits are then

\[ m \pm t s_m \]

where, \( t = 1.96 \) if there is no significant evidence of heterogeneity. Here we have

\[ 1.37 \pm 1.96 \times 0.025 = 1.37 \pm 0.05 \]

giving 1.42 to 1.32 as the required limits.

The value of \( m \) and its fiducial limits should be expressed in actual concentrations, so we find their antilogs and divide by 100 to get back to the original scale, we get
LD 50 = 0.234

With fiducial limits 0.263 and 0.209

A similar calculation gives the fiducial limits of the LD 90.

\[ V_{LD_{90}} = \frac{1}{2.64^2} \left[ \frac{1}{310.46} + \frac{(1.86-1.56)^2}{31.5319} \right] = 0.000 \text{ 871 6} \]

\[ S_{LD_{90}} = \sqrt{0.000 \text{ 871 6}} = 0.0295 \]

Then

\[ 1.86 + 1.96 \times 0.03 = 1.86 + 0.06 \]

giving 1.92 to 1.80 as the fiducial limits. The corresponding doses are

L 90 = 0.724

with fiducial limits 0.832 and 0.631.

Two points should be noted here in connection with fiducial limits. In the first place, formula (7) is merely a close approximation; and in the second place a correction for heterogeneity must be made whenever \( \chi^2 \) is significant. The approximate formula is sufficiently accurate for this example, and also there is no evidence of heterogeneity, so further corrections will not be made. The exact formula for fiducial limits and the use of heterogeneity factor will be demonstrated in the next example.

Table -1 Data from an experiment on the effect of different concentrations of nicotine sulphate on *Drosophila melanogaster*. Determination of empirical probits.

<table>
<thead>
<tr>
<th>(1) Nicotine sulphate gm/100 cc</th>
<th>(2) Number of insects</th>
<th>(3) Number killed</th>
<th>(4) % kill</th>
<th>(5) Log$_{10}$ (Conc. X 100)</th>
<th>(6) Empirical probit</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>137</td>
<td>3</td>
<td>16.8</td>
<td>1.00</td>
<td>4.04</td>
</tr>
<tr>
<td>0.30</td>
<td>152</td>
<td>95</td>
<td>62.5</td>
<td>1.48</td>
<td>5.32</td>
</tr>
<tr>
<td>0.50</td>
<td>146</td>
<td>119</td>
<td>81.5</td>
<td>1.70</td>
<td>5.90</td>
</tr>
<tr>
<td>0.70</td>
<td>154</td>
<td>141</td>
<td>91.6</td>
<td>1.85</td>
<td>6.38</td>
</tr>
<tr>
<td>0.95</td>
<td>152</td>
<td>144</td>
<td>94.7</td>
<td>1.98</td>
<td>6.62</td>
</tr>
</tbody>
</table>
Table -2 Calculation of $\chi^2$ Data of Table -1

<table>
<thead>
<tr>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>Y</td>
<td>P</td>
<td>n</td>
<td>r</td>
<td>nP</td>
<td>r - nP</td>
<td>$\frac{(r - nP)^2}{nP(1 - P)}$</td>
</tr>
<tr>
<td>1.00</td>
<td>4.02</td>
<td>16.4</td>
<td>137</td>
<td>23</td>
<td>22.5</td>
<td>+0.5</td>
<td>0.013</td>
</tr>
<tr>
<td>1.48</td>
<td>5.29</td>
<td>61.4</td>
<td>152</td>
<td>95</td>
<td>93.3</td>
<td>+1.7</td>
<td>0.080</td>
</tr>
<tr>
<td>1.70</td>
<td>5.87</td>
<td>80.8</td>
<td>146</td>
<td>119</td>
<td>118.0</td>
<td>+1.0</td>
<td>0.044</td>
</tr>
<tr>
<td>1.85</td>
<td>6.26</td>
<td>89.6</td>
<td>154</td>
<td>141</td>
<td>138.0</td>
<td>+3.0</td>
<td>0.627</td>
</tr>
<tr>
<td>1.98</td>
<td>6.61</td>
<td>94.6</td>
<td>152</td>
<td>144</td>
<td>143.8</td>
<td>+0.2</td>
<td>0.005</td>
</tr>
</tbody>
</table>

$\chi^2 = 0.769$

Table -3 Calculations – Data of Table -1

<table>
<thead>
<tr>
<th>X</th>
<th>n</th>
<th>Y</th>
<th>w</th>
<th>nw</th>
<th>nwX</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>137</td>
<td>4.0</td>
<td>0.439</td>
<td>60.14</td>
<td>60.14</td>
</tr>
<tr>
<td>1.48</td>
<td>152</td>
<td>5.3</td>
<td>0.616</td>
<td>93.63</td>
<td>138.57</td>
</tr>
<tr>
<td>1.70</td>
<td>146</td>
<td>5.9</td>
<td>0.471</td>
<td>68.77</td>
<td>116.91</td>
</tr>
<tr>
<td>1.85</td>
<td>154</td>
<td>6.3</td>
<td>0.336</td>
<td>51.74</td>
<td>95.72</td>
</tr>
<tr>
<td>1.98</td>
<td>152</td>
<td>6.6</td>
<td>0.238</td>
<td>36.18</td>
<td>71.64</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td>310.46</td>
<td>482.98</td>
</tr>
</tbody>
</table>

$\bar{X} = \frac{482.98}{310.46} = 1.5557$

$\Sigma(nwX^2) = 782.8998$

C.T. = 751.3679

$\Sigma(nwx^2) = 31.5319$
Example - Fitting a probit regression line by the method of maximum likelihood.

The data of columns (1) to (5) of Table -4 were obtained from Dr. W.S. McLeod, being part of a published study on refinements in the technique of testing insecticides. They represent percentage kill in batches of Drosophila treated with nicotine sulphate. The calculations for fitting a probit regression line by the method of maximum likelihood are enumerated below. It is assumed that the student is familiar with the methods of Example -1.

1. Enter data as in column (1) to (5) of Table -4. Obtain empirical probits from the tables by Fisher and Yates, and enter in column (6).
2. Draw Figure -3, showing empirical probits plotted against X. Set up a straight line to give as good an eye-fit as possible. Then read off \( Y_0 \) on the probit scale to one decimal place and enter in column (7).
3. Obtain weighing coefficients from Fisher and Yate’s Table XI or from Finney’s Table II.
4. Calculate and enter \( nw \) in column (9).
5. Calculate working probits from Fisher and Yates’ Table XI and enter in column (10). (Finney’s Table IV is much more rapid than Fisher and Yate’s Table XI and if available should be used). The equation for the working probit is

\[
Y_1 = \left(Y_0 - \frac{P}{Z}\right) + \frac{P}{Z}
\]

Therefore for \( Y_0 = 4.2 \) and \( p = 0.233 \)

\[
Y_1 = 3.4687 + 0.233 \times 3.4519 = 4.27
\]

It will be noted that the working probits in this example are identical with the empirical probits. This is not always the case, although it is to be expected that they will be closer to the empirical probits than to \( Y_0 \).

6. Calculate the entries for columns (11) and (12), and sum columns (9), (11) and (12). From these calculate

\[
\bar{X} = \frac{\sum (nwX)}{\sum (nw)} \quad \bar{Y} = \frac{\sum (nwY_1)}{\sum (nw)}
\]
7. Calculate

\[ \sum (nwx^2) = \sum (nwx^2) - \frac{[\sum (nwx)]^2}{\sum (nw)} \]

\[ \sum (nxwy) = \sum (nxwy) - \frac{[\sum (nxwy)]}{\sum (nw)} \]

\[ \sum (nwy^2) = \sum (nwy^2) - \frac{[\sum (nwy)]^2}{\sum (nw)} \]

\[ \chi^2 = \sum (nwy^2) - \frac{[\sum (nxwy)]^2}{\sum (nwx^2)} \]

In this example \( \chi^2 = 24.91 \), and on referring to Table A-4 we note that this is a significant value for 4 degrees of freedom. Heterogeneity is obvious.

8. Compute

\[ b = \frac{\sum (nxwy)}{\sum (nwx^2)} = \frac{312.86}{118.31} = 2.644 \]

And if heterogeneity were absent we would have

\[ V_b = \frac{1}{\sum (nwx^2)} = \frac{1}{118.31} = 0.008452 \]

\[ S_b = \sqrt{0.008452} = 0.0919 \]
but, heterogeneity being significant, we have

\[ \mu = \frac{\chi^2}{DF} = \frac{24.91}{4} = 6.228 \]

Therefore the corrected variance is

\[ 0.008452 \times 6.228 = 0.052639 \]

and

\[ S_b = \sqrt{0.052639} = 0.2294 \]

Since \( t = 2.78 \) at the 5% point for 4 degrees of freedom, the fiducial limits of \( b \) are obtained from

\[ 2.644 \pm 2.7 \times 0.2294 = 2.644 \pm 0.638 \]

giving limits of 2.01 and 3.28.

9. From the regression equation

\[ Y = \bar{Y} + b(m - \bar{X}) \]

putting \( Y = 5 \), we get

\[ m = \frac{5 - \bar{Y}}{b} + \bar{X} = \frac{5 - 5.0661}{2.644} + 1.0273 = 1.0023 \]

The dose in original units of concentrations expected to give a 50% kill is therefore (antilog 1.0023)/10 = 1.005

Since \( \chi^2 \) indicates significant heterogeneity, it is necessary to take this into account in determining the fiducial limits. Also in this example we shall apply the formula for the more exact fiducial limits.
First we have $\mu$, as calculated in step 8, which must be introduced in order to take care of heterogeneity. Also we must obtain

$$g = \frac{t^2 \mu}{b^2 \sum (nw^2)} = \frac{2.776^2 \times 6.228}{2.644^2 \times 118.31} = 0.0580$$

Then the exact fiducial limits are given by

$$m + \frac{g}{1 - g} (m - \bar{X}) \pm \frac{t}{b(1-g)} \sqrt{\frac{1 - g}{\sum (nw)} + \frac{(m - \bar{X})^2}{\sum (nw^2)}} \mu$$

Which for purposes of calculation is most conveniently broken down into three parts.

$$m + \frac{g}{1 - g} (m - \bar{X}) = 1.0023 + \frac{0.0580}{0.9420} (1.0023 - 1.0273) = 1.0008$$

$$\frac{t}{b(1-g)} = \frac{2.776}{2.644 \times 0.9420} = 1.1146$$

$$\sqrt{\frac{\mu}{\sum (nw)} + \frac{(m - \bar{X})^2}{\sum (nw^2)}} = \sqrt{\frac{0.9420}{3827} + \frac{(1.0023 - 10273)^2}{118.31}} = 6.228 \approx 0.03953$$

Finally, $1.008 \pm 1.1146 \times 0.03953 = 1.008 \pm 0.0441$

giving fiducial limits of 1.045 to 0.957 in dosage terms and 1.11 to 0.91 in actual concentrations.

The procedure and methods for calculating fiducial limits of $m$ may be summarized as follows:

a. No heterogeneity ($\chi^2$ not significant)
(i) \( g \) small with respect to unity, where

\[
g = \frac{t^2 \mu}{b^2 \sum (nw)^2} \quad (taking \ t = 1.96)
\]

Limits given by

\[
m \pm 1.96 \sqrt{\frac{1}{b^2} \left[ \frac{1}{\sum (nw)} + \frac{(m - \bar{X})^2}{\sum (nw)^2} \right]}
\]

(ii) \( g \) appreciable with respect to unity.

Limits given by

\[
m + \frac{g}{1-g} (m - \bar{X}) \pm \frac{t}{b(1-g)} \sqrt{\frac{1-g}{\sum (nw)} + \frac{(m - \bar{X})^2}{\sum (nw)^2}}
\]

b. Significantly heterogeneity

Calculate

\[
\mu = \frac{\chi^2}{DF} \quad and \quad g = \frac{t^2 \mu}{b^2 \sum (nm)^2}
\]

\[
m + \frac{g}{1-g} (m - \bar{X}) \pm \frac{t}{b(1-g)} \sqrt{\frac{1-g}{\sum (nw)} + \frac{(m - \bar{X})^2}{\sum (nw)^2} \mu}
\]

At the point the student might refer to experimental design with a view of assessing the design of the experiment just analyzed.
Fig – 3 Fitting of probit regression line to data of Table -4.

It would be noted that the dosage range is very good in that the percentage kills range from 23 to 75%. Also, the number of insects in each batch is quite satisfactory. In this experiment there is no control level and hence no measure of natural mortality. This might be considered a weakness in the design, but it is of course possible that the experimenter may have had natural mortality in good control and considered that a measure of it would not required. A significant feature of this experiment is the heterogeneity, and an examination of Figure -3 indicates a curvilinear trend. Possibly some relation between dose and dosage other than the logarithmic one tried would be better, and this point at least requires investigation.

**Example -. Determination of relative potency of biological preparations.**

In the measurement of the effect of toxic substances, effectiveness of sera for the prevention of disease, and so forth, one of the common problems is to comparing the potency of two or more preparations. In some experiments the preparations may be new ones, but the usual problem is that of comparing one or more new preparations with a standard. In biological work of this sort it is very difficult to obtain exactly similar conditions in experiments carried out at different times and with different insects or animals. The inclusion of a standard in an experiment is therefore a routine procedure.
Irwin quotes an example from Smith in which two anti-pneumococcus sera are tested for the prevention of disease in mice. The data are given in Table -5. The symbols α and β distinguish the sera, and form our purpose β can be considered as the standard. Note that the data are for percentage survival and not for percentage mortality as in previous examples. The calculations are conveniently carried out in the steps enumerated below:

1. Set up the data in the first four columns of Table -3 and calculate p (the percentage survival). Empirical probits are then found from Fisher and Yates’ Table IX, noting that for p = 0 the empirical probit is -∞.

2. Draw figure -4, locating the points corresponding to the empirical probits but neglecting -∞. On this graph draw two parallel lines giving a good eye fit. Then for each point on the graph find the corresponding point on the straight line and

3. 

Figure 5 : Probit regression line of the data given in Table 5

4. from these read off the provisional probits (Y₀) on the probit scale. Enter these in column (7). From the empirical probit -∞, note that the figure in entered at X = 0.64 for α and X = 1.24 for β.

5. Enter weighting coefficient in column (8) from Fisher and Yates’ Table XI. Calculate and enter nw in column 9.

6. The working probits of column (10) are obtained from Fisher and Yates’ Table IX. For Y₀ = 2.7, Y₁ = 2.32 from the formula for a minimum working probit; that is, Y₁ = Y - P / Z. For Y₀ = 3.6 we have
\[
Y_1 = \left( Y - \frac{P}{Z} \right) + \frac{P}{Z} = 3.0606 + 0.05 \times 6.6788 = 3.39
\]

5. Calculate and enter \( nw \) \( X \) and \( nw \) \( Y \) in columns (11) and (12). Then sum columns (9), (11), (12) separately for \( \alpha \) and \( \beta \), and calculate \( 1/\sum(nw) \), \( \bar{X} \), and \( \bar{Y} \).

6. Provisional values of \( b_1 \), \( m_\alpha \), \( m_\beta \), \( M_{\alpha \beta} \) and \( \rho_{\alpha \beta} \) may be calculated.

\[
M_{\alpha \beta} = m_\beta - m_\alpha \text{ (relative potency in dosage)}
\]

\[
\rho_{\alpha \beta} = 10^{M_{\alpha \beta}} \text{ (relative potency in unit of concentration)}
\]

From the graph for \( \alpha \)

\[
b = \frac{3.89}{1.35} = 2.88
\]

\[
X_2 = 1.80 \quad Y_2 = 6.01
\]

\[
X_1 = 0.45 \quad Y_1 = 2.12
\]

Difference = 1.35 \quad 3.89

From the graph for \( \beta \)

\[
b = \frac{3.60}{5.70_{1.25}} = 2.88
\]

\[
X_2 = 2.40 \quad Y_2 = 5.70
\]

\[
X_1 = 1.15 \quad Y_1 = 2.10
\]

Difference = 1.25 \quad 3.60

The two estimates of \( b \) must of course agree if the two lines have been drawn parallel.

\[Y = \bar{Y} + b(X - \bar{X})\]

Then setting up the regression equation for both \( \alpha \) and \( \beta \) and substituting \( m_\alpha \) \( m_\beta \) for \( X \) and 5 for \( Y \) gives
\[ \alpha \quad Y = 4.96 + 2.88 (X - 1.39) = 0.96 + 2.88 X \\
5 = 0.96 + 2.88 \frac{m_\alpha}{2.88} = 1.40 \]
\[ m_\alpha = \frac{4.04}{2.88} \]

\[ \square \quad Y = 4.87 + 2.88 (X - 2.08) = -1.12 + 2.88 X \\
5 = -1.12 + 2.88 \frac{m_\beta}{2.88} = 2.12 \]
\[ m_\beta = \frac{6.12}{2.88} \]

\[ M_\alpha = 2.12 - 1.40 = 0.72 \]
\[ \rho_\alpha = 10^{0.72} = 5.2 \]

The provisional estimate indicates that 5.2 times the concentration of \( \square \) is required to give the same protection a 1 unit of concentration of \( \alpha \).

7. Complete the calculations given below in order to obtain new estimates of the parameters, and to test their significance and the goodness of fit of the new regression lines.

<table>
<thead>
<tr>
<th></th>
<th>( \alpha )</th>
<th>( \square )</th>
<th>( \alpha + \square )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Sigma (nwX^2) )</td>
<td>160.1200</td>
<td>322.5672</td>
<td></td>
</tr>
<tr>
<td>CT</td>
<td>151.4806</td>
<td>315.4643</td>
<td></td>
</tr>
<tr>
<td>( \Sigma (nwX^2) )</td>
<td>8.6394+</td>
<td>7.1029 = 15.7423</td>
<td></td>
</tr>
<tr>
<td>( \Sigma (nwXY) )</td>
<td>566.2667</td>
<td>756.1954</td>
<td></td>
</tr>
<tr>
<td>CT</td>
<td>540.0853</td>
<td>739.8953</td>
<td></td>
</tr>
<tr>
<td>( \Sigma (nwxY) )</td>
<td>26.1814+</td>
<td>16.3001= 42.4815</td>
<td></td>
</tr>
<tr>
<td>( \Sigma (nwY^2) )</td>
<td>2009.2773</td>
<td>1777.4078</td>
<td></td>
</tr>
<tr>
<td>CT</td>
<td>1925.6069</td>
<td>1735.3626</td>
<td></td>
</tr>
<tr>
<td></td>
<td>83.6704</td>
<td>42.0452= 125.7156</td>
<td></td>
</tr>
<tr>
<td>Regression</td>
<td>7.3418</td>
<td>37.4067</td>
<td></td>
</tr>
<tr>
<td>( \chi^2 )</td>
<td>4.3286 +</td>
<td>4.6389 = 8.9675</td>
<td></td>
</tr>
<tr>
<td>DF</td>
<td>3 +</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>5% point</td>
<td>7.815</td>
<td>7.815</td>
<td>12.592</td>
</tr>
</tbody>
</table>
A complete analysis of $\chi^2$ can be made by calculating a total $\chi^2$ for degrees of freedom as follows:

$$\chi^2 = 125.7156 - \frac{(42.415)^2}{15.7423} = 11.0768$$

giving the analysis

<table>
<thead>
<tr>
<th></th>
<th>$\chi^2$</th>
<th>DF</th>
<th>MS</th>
<th>5% point $\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parallelism of regression</td>
<td>2.1093</td>
<td>1</td>
<td>2.109</td>
<td>3.84</td>
</tr>
<tr>
<td>Heterogeneity</td>
<td>8.9675</td>
<td>6</td>
<td>1.494</td>
<td>12.59</td>
</tr>
<tr>
<td>Total</td>
<td>11.0768</td>
<td>7</td>
<td>1.582</td>
<td>14.07</td>
</tr>
</tbody>
</table>

None of the $\chi^2$ values is significant at the 5% point although there is some indication of heterogeneity. This seems to be due to a tendency of the $\beta$ regression to be somewhat curvilinear and to one point on the $\beta$ regression that deviates rather widely from the straight line. The parallelism is also only reasonably satisfactory. Note that heterogeneity as measured here is a residual effect that is, it represents total lack of agreement with the straight line within the two regressions.

Due to some indication of heterogeneity the heterogeneity factor will be used in setting up fiducial limits although it is not expected that they will be greatly affected.

8. Obtain new estimates of the parameters and set up fiducial limits.

$$\mu = \frac{8.9675}{4} = 2.2419 \quad b = \frac{42.4815}{15.7423} = 2.6986$$

$$S_b = \sqrt{0.1424} = 0.377$$

Fiducial limits are $2.699 + 2.776 \times 0.377$, giving 3.75 and 1.65. Then from
\[ Y_\alpha = 4.9597 + 2.6986 (X - 1.3911) = 1.2057 + 2.6986X \]

The value of \( m_\alpha \) is calculated by putting \( m_\alpha = X \) and \( Y_\alpha = 5 \).

\[
m_\alpha = \frac{5 - 1.2057}{2.6986} = 1.4060
\]

Similarly

\[ Y_\beta = 4.8677 + 2.6986 (X - 2.0754) = -0.7330 + 2.6986X \]

\[
m_\beta = \frac{5 + 0.7330}{2.6986} = 2.1244
\]

Then

\[ M_{\alpha\beta} = 2.1244 \cdot 1.4060 = 0.7184 \]

\[ \rho_{\alpha\beta} = 10^{0.72} = 5.2 \]

The fiducial limits of \( M_{\alpha\beta} \) are given by

\[
M + \frac{g}{1-g} [M - (\bar{X}_\beta - \bar{X}_\alpha)] \pm \frac{t}{b(1-g)} \sqrt{(1-g)(V_{\bar{Y}_\alpha} + V_{\bar{Y}_\beta}) + [(\bar{X}_\beta - \bar{X}_\alpha) - M]^2 V_b}
\]

Where, \( g \) and the variances must in this example, be increased by the heterogeneity factor. We have to find:

\[
g = \frac{2.776^2 \times 2.2419}{2.6986^2 \times 15.7423} = \frac{17.2765}{114.6424} = 0.1507
\]

\[
V_{\bar{Y}_\alpha} = \frac{2.2419}{78.28} = 0.02864
\]
\[ V_{\bar{\beta}} = \frac{2.2419}{73.24} = 0.03061 \]

\[ V_b = 0.1424 \text{ (step 8, above)} \]

\[ \frac{g}{1-g} \left[ M - \left( \bar{X}_\beta - \bar{X}_\alpha \right) \right] = \frac{0.1507}{0.8493} (0.7184 - 0.6843) = 0.006051 \]

\[ \frac{t}{b(1-g)} = \frac{2.776}{2.6986 \times 0.8493} = 1.2112 \]

\[ \sqrt{(1-g) \left( V_{\bar{\beta}} + V_{\bar{\alpha}} \right) + \left[ \left( \bar{X}_\beta - \bar{X}_\alpha \right) - M \right]^2 V_b} \]

\[ \sqrt{(0.8493 \times 0.05925) + (0.6843 - 0.7184)^2 \times 0.1424} \]

\[ = \sqrt{0.05049} = 0.2247 \]

Finally the fiducial limits are:

\[ 0.7184 + 0.0061 \pm 1.2112 \times 0.2247 = 0.7245 \pm 0.2722 \]

giving 0.9967 and 0.4523. The corresponding relative potencies are \(10^{0.997} = 9.93\) and \(10^{0.452} = 2.83\)

owing largely to lack of homogeneity the final result in terms of relative potency of the two preparations is not very satisfactory.
Table -5

Calculation of relative potency of two anti-pneumococcus sera

<table>
<thead>
<tr>
<th></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>(11)</th>
<th>(12)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\log_{10} x$</td>
<td>Number surviving</td>
<td>Empirical point</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda$</td>
<td>0.000 4375</td>
<td>0.64</td>
<td>40</td>
<td>0</td>
<td>0</td>
<td>$\infty$</td>
<td>2.7</td>
<td>0.076</td>
<td>3.04</td>
<td>2.32</td>
<td>1.946</td>
<td>7.053</td>
</tr>
<tr>
<td></td>
<td>0.000 875</td>
<td>0.94</td>
<td>40</td>
<td>2</td>
<td>5</td>
<td>3.36</td>
<td>3.6</td>
<td>0.302</td>
<td>12.08</td>
<td>3.39</td>
<td>11.355</td>
<td>40.951</td>
</tr>
<tr>
<td></td>
<td>0.001 75</td>
<td>1.24</td>
<td>40</td>
<td>14</td>
<td>35</td>
<td>4.61</td>
<td>4.4</td>
<td>0.558</td>
<td>22.32</td>
<td>4.63</td>
<td>27.677</td>
<td>103.342</td>
</tr>
<tr>
<td></td>
<td>0.003 5</td>
<td>1.54</td>
<td>40</td>
<td>30</td>
<td>75</td>
<td>5.67</td>
<td>5.3</td>
<td>0.616</td>
<td>24.64</td>
<td>5.65</td>
<td>37.946</td>
<td>139.216</td>
</tr>
<tr>
<td></td>
<td>0.007</td>
<td>1.85</td>
<td>40</td>
<td>34</td>
<td>85</td>
<td>6.04</td>
<td>6.1</td>
<td>0.405</td>
<td>16.20</td>
<td>6.03</td>
<td>29.970</td>
<td>97.686</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.001 75</td>
<td>1.25</td>
<td>40</td>
<td>0</td>
<td>0</td>
<td>$\infty$</td>
<td>2.4</td>
<td>0.040</td>
<td>1.60</td>
<td>2.06</td>
<td>1.984</td>
<td>3.296</td>
</tr>
<tr>
<td></td>
<td>0.003 5</td>
<td>1.54</td>
<td>40</td>
<td>2</td>
<td>5</td>
<td>3.36</td>
<td>3.2</td>
<td>0.180</td>
<td>7.20</td>
<td>3.38</td>
<td>11.088</td>
<td>24.336</td>
</tr>
<tr>
<td></td>
<td>0.007</td>
<td>1.85</td>
<td>40</td>
<td>14</td>
<td>35</td>
<td>4.61</td>
<td>4.1</td>
<td>0.471</td>
<td>18.84</td>
<td>4.72</td>
<td>34.854</td>
<td>88.925</td>
</tr>
<tr>
<td></td>
<td>0.014</td>
<td>2.15</td>
<td>40</td>
<td>19</td>
<td>47.5</td>
<td>4.94</td>
<td>5.0</td>
<td>0.637</td>
<td>25.48</td>
<td>4.94</td>
<td>54.782</td>
<td>125.871</td>
</tr>
<tr>
<td></td>
<td>0.028</td>
<td>2.45</td>
<td>40</td>
<td>30</td>
<td>75</td>
<td>5.67</td>
<td>5.8</td>
<td>0.503</td>
<td>20.12</td>
<td>5.67</td>
<td>49.294</td>
<td>114.080</td>
</tr>
<tr>
<td></td>
<td>73.24</td>
<td>132.002</td>
<td>356.080</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
10. **Experimental design:**

Two questions normally arise in the mind of investigators planning experiments of this type. In the first place they ask: “What should be the range and number of concentrations?” and in the second place: “How many insects or animals, etc. should be treated at each level?”

The range of dose is obviously something that cannot be answered without some preliminary information, and in dealing with an entirely new substance it would seem to be desirable to get preliminary data on roughly approximate level for the LD 10, LD 50 an LD 90. It should then be possible to set up a series of satisfactory dose levels for a well controlled experiment, keeping in mind that the least information with respect to potency is given by the levels that give very low and very high percentage kills. The most valuable points are those giving kills from 25 to 75%. Kills lower than 16% and higher than 84% give so little information with respect to LD 50 that they can practically can be disregarded. Their precision is so low that a much larger number of insects is required to estimate them with the accuracy ordinarily obtained in the region of LD 50.

The question of how many insects there should be in each batch is related to the number of batches, but in any event it is preferable with a given number of insects available to spread these over several batches at different levels rather than to have only 3 to 4 larger batches at 3 or 4 levels. The general accuracy of experiment is of course increased by increasing the number of insects in each batch. Perhaps certain published results of probit analysis with 20 to 30 insects in each batch have led investigators to think that a greater number is unnecessary. This is of course incorrect, and the greatest number should be used that does not make the experiment too expensive and unwieldy relative to the value of the information to be obtained. It must be recognized that in experimenting with animals the numbers may have to be small. The chief point to remember, however, is that the probit method or any other type of analysis does not make up for lack of accuracy in the experiment or unsatisfactory techniques.

In most experiments of the biological assay type it is important to have some measure of natural mortality. This is usually accomplished by having a control level. When mortality occurs at this level it is advisable to make a correction in the analysis.
References

CHAPTER 13

Multivariate Techniques

1. Introduction
The researchers in biological, physical and social sciences frequently collect measurements on several variables. Generally the data is analyzed by taking one variable at a time. The inferences drawn by analyzing the data for each of the variables may be misleading. This can best be explained from the story of the six blind persons, who tried to describe an elephant after each one touching and feeling a part of it. All of us know that they came out with six different versions of what an elephant was like, each version being partially correct but none was near to reality. Therefore, the data on several variables should be analyzed using multivariate analytical techniques.

Various statistical methods for describing and analyzing these multivariate data sets are Hotelling $T^2$; Multivariate analysis of variance (MANOVA), Discriminant Analysis, Principal Component Analysis, Factor Analysis, Canonical Correlation Analysis, Cluster Analysis, etc. In this talk, we present an overview of the multivariate analytical techniques.

1. Testing of mean vector - One Sample Case
This is useful for the situations where the data on the different variables are collected and it is required to test whether the sample mean vectors is equal to a specified mean vector. To be specific: Let $x_1, x_2, ..., x_n$ be a random sample of size $n$ is drawn from the population with $p$-dimensional mean vector $\mu_0$ and based on this sample we want to test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$.

If variance covariance matrix $\Sigma$ is known or the sample is large, $\chi^2$ test is used.

$$\chi^2 = n(\bar{x} - \mu_0)' \Sigma^{-1} (\bar{x} - \mu_0)$$

with $p$ degrees of freedom where $\bar{x} = \frac{1}{n} \sum_{j=1}^{n} x_j$ is the sample mean calculated from the sample, $p$ is number of variable in the study.
If \( \Sigma \) is not known and sample size is small, Hotelling \( T^2 \) is used.

\[
T^2 = n(\bar{x} - \mu_0)'s^{-1}(\bar{x} - \mu_0)
\]

where \( \bar{x} = \frac{1}{n} \sum_{j=1}^{n} x_j \), \( s = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x})(x_j - \bar{x})' \)

\[
\frac{(n-p)T^2}{(n-1)p} \approx F_{p, n-p}
\]

**Example 1:** (Example 5.2 in Johnson and Wichern, 2002). Perspiration from 20 healthy females was analyzed. Three components, \( X_1 = \) sweat rate, \( X_2 = \) sodium content and \( X_3 = \) potassium content were measured and the results are presented in table 1.

<table>
<thead>
<tr>
<th>Individual</th>
<th>( X_1 ) (sweat rate)</th>
<th>( X_2 ) (sodium content)</th>
<th>( X_3 ) (potassium content)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.7</td>
<td>48.5</td>
<td>9.3</td>
</tr>
<tr>
<td>2</td>
<td>5.7</td>
<td>65.1</td>
<td>8.0</td>
</tr>
<tr>
<td>3</td>
<td>3.8</td>
<td>47.2</td>
<td>10.9</td>
</tr>
<tr>
<td>4</td>
<td>3.2</td>
<td>53.2</td>
<td>12.0</td>
</tr>
<tr>
<td>5</td>
<td>3.1</td>
<td>55.5</td>
<td>9.7</td>
</tr>
<tr>
<td>6</td>
<td>4.6</td>
<td>36.1</td>
<td>7.9</td>
</tr>
<tr>
<td>7</td>
<td>2.4</td>
<td>24.8</td>
<td>14.0</td>
</tr>
<tr>
<td>8</td>
<td>7.2</td>
<td>33.1</td>
<td>7.6</td>
</tr>
<tr>
<td>9</td>
<td>6.7</td>
<td>47.4</td>
<td>8.5</td>
</tr>
<tr>
<td>10</td>
<td>5.4</td>
<td>54.1</td>
<td>11.3</td>
</tr>
<tr>
<td>11</td>
<td>3.9</td>
<td>36.9</td>
<td>12.7</td>
</tr>
<tr>
<td>12</td>
<td>4.5</td>
<td>58.8</td>
<td>12.3</td>
</tr>
<tr>
<td>13</td>
<td>3.5</td>
<td>27.8</td>
<td>9.8</td>
</tr>
<tr>
<td>14</td>
<td>4.5</td>
<td>40.2</td>
<td>8.4</td>
</tr>
<tr>
<td>15</td>
<td>1.5</td>
<td>13.5</td>
<td>10.1</td>
</tr>
<tr>
<td>16</td>
<td>8.5</td>
<td>56.4</td>
<td>7.1</td>
</tr>
<tr>
<td>17</td>
<td>4.5</td>
<td>71.6</td>
<td>8.2</td>
</tr>
<tr>
<td>18</td>
<td>6.5</td>
<td>52.8</td>
<td>10.9</td>
</tr>
<tr>
<td>19</td>
<td>4.1</td>
<td>44.1</td>
<td>11.2</td>
</tr>
<tr>
<td>20</td>
<td>5.5</td>
<td>40.9</td>
<td>9.4</td>
</tr>
</tbody>
</table>
Test the hypothesis, \( H_0 : \mu = \mu_0 \) given \( \mu_0 = \begin{bmatrix} 4 & 50 & 10 \end{bmatrix} \) against \( H_1 : \mu \neq \mu_0 \). From Table 1, we can calculate
\[
\bar{x} = \frac{1}{n} \sum_{j=1}^{n} X_j = \begin{bmatrix} 4.640 \\ 45.400 \\ 9.965 \end{bmatrix}, \quad s = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x})(x_j - \bar{x}) = \begin{bmatrix} 2.879368 & 10.01 & -1.89095 \\ 10.01 & 199.7884 & -5.64 \\ -1.89095 & -5.64 & 3.627658 \end{bmatrix}
\]
and the observed \( T^2 \) value is
\[
= n(\bar{x} - \mu_0)'s^{-1}(\bar{x} - \mu_0) = 20 \begin{bmatrix} 4.640 - 4 \\ 45.400 - 50 \\ 9.965 - 10 \end{bmatrix}^{-1} \begin{bmatrix} 2.879368 \\ 10.01 \\ -1.89095 \end{bmatrix} = 9.738774
\]
Comparing the observed \( T^2 = 9.738774 \) with the critical value
\[
\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha) = 3.353 \times 3.20 = 10.73 \text{ we see that } T^2 = 9.74 < 10.73, \text{ and consequently we accept } H_0.
\]

2. Testing of mean vectors - Two Sample Case
Consider that we have two independent random samples of sizes \( n_1 \) and \( n_2 \) with mean vectors \( \bar{x}_1 \) and \( \bar{x}_2 \) and sample dispersion matrices \( s_1 \) and \( s_2 \) respectively and want to test the hypothesis
\( H_0 : \mu_1 = \mu_2 \) against \( H_1 : \mu_1 \neq \mu_2 \)
\( \mu_1 \) and \( \mu_2 \) are mean vectors of populations from which samples are drawn. If population dispersion matrices are unknown but same, we use
\[
T^2 = \frac{n_1n_2}{n_1 + n_2} (\bar{x}_1 - \bar{x}_2)'s^{-1}_\text{pooled}(\bar{x}_1 - \bar{x}_2)
\]
where \( s^{-1}_\text{pooled} = \frac{(n_1-1)s_1 + (n_2-1)s_2}{n_1 + n_2 - 2} \).
\( T^2 \) is distributed as \( \frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)}F_{p,n_1+n_2-p-1} \)
Example 2: (Example 6.4 in Johnson and Wichern, 2002). Samples of sizes $n_1 = 45$ and $n_2 = 55$ were taken of homeowners with and without air conditioning respectively. Two measurements of electrical usage (in kilowatt-hours) were considered. The first is a measure of total on-peak consumption ($x_1$) during July and the second is a measure of total off-peak consumption during July. Test whether there is a difference in electrical consumption between those with air conditioning and those without.

The summary statistics given are

$$
\begin{align*}
\bar{x}_1 &= \begin{bmatrix} 204.4 \\ 556.6 \end{bmatrix} & \bar{x}_2 &= \begin{bmatrix} 130.0 \\ 355.0 \end{bmatrix} \\
S_1 &= \begin{bmatrix} 13825.3 & 23823.4 \\ 23823.4 & 73107.4 \end{bmatrix} & S_2 &= \begin{bmatrix} 8632.0 & 19616.7 \\ 19616.7 & 55964.5 \end{bmatrix}
\end{align*}
$$

$n_1 = 45, n_2 = 55$

Here the null hypothesis is $H_0: \mu_1 = \mu_2$ and alternate hypothesis is $H_1: \mu_1 \neq \mu_2$. To test the difference, first we calculate

$$s_{pooled} = \frac{(n_1 - 1)s_1 + (n_2 - 1)s_2}{n_1 + n_2 - 2}$$

$$= \begin{bmatrix} 10963.7 & 21505.5 \\ 21505.5 & 63661.3 \end{bmatrix}$$

Now

$$T^2 = \frac{n_1n_2}{n_1 + n_2} (\bar{x}_1 - \bar{x}_2)^T s_{pooled}^{-1} (\bar{x}_1 - \bar{x}_2)$$

$$= 16.22066$$

Comparing the observed $T^2$ with the critical value

$$\frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F_{p, n_1 + n_2 - p - 1}(\alpha) = \frac{98(2)}{97} F_{2.97(0.05)} = 6.26$$

We see that the observed $T^2 = 16.22066 > 6.26$, we reject the null hypothesis and conclude that there is a difference in electrical consumption between those with air conditioning and those without.

Note:

(i) For this testing, Mahalanobis $D^2$ can also be used which is a linear function of $T^2$

$$D^2 = (\bar{x}_1 - \bar{x}_2)^T s_{pooled}^{-1} (\bar{x}_1 - \bar{x}_2)$$

$$= \frac{n_1n_2}{n_1 + n_2} T^2$$
(ii) If $\Sigma_1 \neq \Sigma_2$, the above test cannot be used. For large sample size or dispersion matrices known, $\chi^2$ test can be used. However, test for small sample sizes and dispersion matrices not known to be equal is beyond the scope of discussion. Readers may go through the references given at the end.

**Steps to Carry out the Analysis: Testing Mean Vector(s) (Using MS-EXCEL)**

We to use the inbuilt Functions of MS-EXCEL like Average; Mean; VAR: Variance and COVAR=m/(n-1): Covariance. Correlation can be obtained using the function CORREL.

**Matrix Inverse**

Mark the area for the resultant matrix → Formula bar → =minverse (mark range of original matrix) → press control + shift + enter

**Matrix multiplication**

Mark the area for the resultant matrix → Formula bar → =mmult (mark range of first matrix, mark range of second matrix) → press control + shift + enter

Using the matrix multiplication and matrix inversion one can easily calculate Hotelling's $T^2$.

3. **Multivariate Analysis of Variance (MANOVA)**

**One way Classified Data**

Consider that the random samples from each of $g$ (say) populations using are arranged as

Population 1: $x_{11}, x_{12}, \ldots, x_{1n_1}$

Population 2: $x_{21}, x_{22}, \ldots, x_{2n_2}$

\[\vdots\]

Population $g$: $x_{g1}, x_{g2}, \ldots, x_{gn_g}$

Multivariate analysis of variance is used first to investigate whether the populations mean vectors are the same and, if not, which mean components differ significantly. MANOVA is carried out under the following two assumptions: 1. Dispersion matrices of various populations are same. 2. Each population is multivariate normal. One-way Classified MANOVA Table for testing the equality of $g$-population mean Vectors is given below:
<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Degrees of freedom</th>
<th>SSP matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population or treatment</td>
<td>g-1</td>
<td>$T = \sum_{i=1}^{g} n_i (\bar{x}_i - \bar{x})(\bar{x}_j - \bar{x})'$</td>
</tr>
<tr>
<td>Residual (error)</td>
<td>$\sum_{i=1}^{g} n_i - g$</td>
<td>$R = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}<em>i)(x</em>{ij} - \bar{x}_i)'$</td>
</tr>
<tr>
<td>Total</td>
<td>$\sum_{i=1}^{g} n_i - l$</td>
<td>$T + R = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}<em>i)(x</em>{ij} - \bar{x})'$</td>
</tr>
</tbody>
</table>

We reject the null hypothesis of equal mean vectors if the ratio of generalized variance (Wilk's lambda statistic) $A^* = \frac{|R|}{|T + R|}$ is too small. The distribution of $A^*$ in different cases are as below.

$p = 1 \quad g \geq 2 \quad \left(\frac{\sum n_i - g}{g - 1}\right)\left(\frac{1 - A^*}{A^*}\right) - F_{g-1, (\sum n_i - g)}(\alpha)$

$p = 2 \quad g \geq 2 \quad \left(\frac{\sum n_i - g - 1}{g - 1}\right)\left(\frac{1 - \sqrt{A^*}}{\sqrt{A^*}}\right) \sim F_{2(g-1), 2(\sum n_i - g - 1)}(\alpha)$

$p \geq 1 \quad g = 2 \quad \left(\frac{\sum n_i - p - 1}{p}\right)\left(\frac{1 - A^*}{A^*}\right) \sim F_{p, (\sum n_i - p - 1)}(\alpha)$

$p \geq 1 \quad g = 3 \quad \left(\frac{\sum n_i - p - 2}{p}\right)\left(\frac{1 - \sqrt{A^*}}{\sqrt{A^*}}\right) \sim F_{2p, 2(\sum n_i - p - 2)}(\alpha)$

and for other cases $-\left(n - 1 - \frac{(p + g)}{2}\right) \ln A^* \sim \chi^2_{p(g-1)}(\alpha)$ (approximate).
Example 3: (Example 6.8 in Johnson and Wichern, 2002). Consider the following independent samples:

<table>
<thead>
<tr>
<th></th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population 1</td>
<td>9</td>
<td>3</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>Population 2</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Population 3</td>
<td>3</td>
<td>8</td>
<td>1</td>
<td>24</td>
</tr>
<tr>
<td>Grand Total</td>
<td>12</td>
<td>14</td>
<td>12</td>
<td>40</td>
</tr>
</tbody>
</table>

Due to variable 1

Sum of squares (Population) = \( \frac{24^2}{3} + \frac{6^2}{3} + \frac{38^2}{8} = 78 \)

Sum of squares (Total) = \( 9^2 + 6^2 + \ldots + 2^2 - \frac{38^2}{8} = 88 \)

Sum of squares (Residual) = 88 – 78 = 10 (by subtraction)

Due to variable 2

Sum of squares (Population) = \( \frac{12^2}{3} + \frac{4^2}{2} + \frac{24^2}{3} - \frac{40^2}{8} = 48 \)

Sum of squares (Total) = \( 3^2 + 2^2 + \ldots + 7^2 - \frac{40^2}{8} = 72 \)

Sum of squares (Residual) = 72 – 48 = 24 (by subtraction)

Due to variable 1 and 2

Sum of cross products (Population) = \( \frac{24 \times 12}{3} + \frac{2 \times 4}{2} + \frac{6 \times 24}{3} - \frac{32 \times 40}{8} = -12 \)

Sum of cross products (Total) = \( 9 \times 3 + 6 \times 2 + \ldots + 2 \times 7 - \frac{32 \times 40}{8} = -11 \)

Sum of cross products (Residual) = –11 – (–12) = 1
### MANOVA

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Degrees of freedom</th>
<th>SSP matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>$3 - 1 = 2$</td>
<td>$\begin{bmatrix} 78 &amp; -12 \ -12 &amp; 48 \end{bmatrix}$</td>
</tr>
<tr>
<td>Residual (error)</td>
<td>$3 + 2 + 3 - 3 = 5$</td>
<td>$\begin{bmatrix} 10 &amp; 1 \ 1 &amp; 24 \end{bmatrix}$</td>
</tr>
<tr>
<td>Total</td>
<td>$3 + 2 + 3 - 1 = 7$</td>
<td>$\begin{bmatrix} 88 &amp; -11 \ -11 &amp; 72 \end{bmatrix}$</td>
</tr>
</tbody>
</table>

To test the hypothesis $H_0: \mu_1 - \mu_2 - \mu_3$. We use Wilk's lambda statistic

$$\Lambda^* = \left| \frac{R}{T + R} \right| = \left| \begin{bmatrix} 10 & 1 \\ 1 & 24 \end{bmatrix} \right| = \frac{10(24) - (1)^2}{88(72) - (-11)^2} = \frac{239}{6215} = 0.0385$$

Since $p = 2$ (variables) and $g = 3$ (populations), we use the following

$$\frac{\sum \eta_j - p - 2}{p} \left( \frac{1 - \sqrt{\Lambda^*}}{\sqrt{\Lambda^*}} \right) = \left( \frac{8 - 3 - 1}{3 - 1} \right) \left( 1 - \sqrt{0.0385} \right)^2 = 8.19$$

with a percentage point of an $F$-distribution having $\eta_1 = 4$ & $\eta_2 = 8$ d.f. Since $8.19 > F_{4,8}(0.01) = 7.01$, we reject the null hypothesis at 1% level of significance and conclude that there exists treatment differences. The pairwise comparisons can be done using the contrast analysis.

### Two way Classified Data

Consider that the data are recorded at various levels of two factors. This situation is identical to that of an experiment conducted using a randomized complete block (RCB) design to compare $v$ treatments with $r$ blocks and the data is collected on $p$-variables. Let $y_{ijk}$ denote the observed value of the $k^{th}$ response variable for the $i^{th}$ treatment in the $j^{th}$ replication

$i = 1,2, \ldots, v; j = 1,2, \ldots, r; k = 1,2, \ldots, p$. For performing the MANOVA, the data from the experimental set up can be rearranged as
Here $y_{ij} = (y_{i1} \ y_{i2} \ \cdots \ y_{ijk} \ \cdots \ y_{ij})^\prime$ is a $p$-variate vector of observations taken from the plot receiving the treatment $i$ in replication $j$. $\bar{y}_i = \frac{1}{r} \sum_{j=1}^{r} y_{ij}$; $\bar{y}_j = \frac{1}{v} \sum_{i=1}^{v} y_{ij}$ and $\bar{y}_{..} = \frac{1}{vr} \sum_{i=1}^{v} \sum_{j=1}^{r} y_{ij}$.

An outline of MANOVA Table for testing the equality of treatment effects and replication effects is

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>SSCPM (Sum of Squares and Cross Product Matrix)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>$v-1 = h$</td>
<td>$H = b \sum_{i=1}^{v} (\bar{y}<em>i - \bar{y}</em>{..})(\bar{y}<em>i - \bar{y}</em>{..})^\prime$</td>
</tr>
<tr>
<td>Replication</td>
<td>$r-1 = t$</td>
<td>$B = v \sum_{j=1}^{b} (\bar{y}<em>j - \bar{y}</em>{..})(\bar{y}<em>j - \bar{y}</em>{..})^\prime$</td>
</tr>
<tr>
<td>Residual</td>
<td>$(v-1)(r-1) = s$</td>
<td>$R = \sum_{i=1}^{v} \sum_{j=1}^{b} (y_{ij} - \bar{y}<em>i - \bar{y}<em>j + \bar{y}</em>{..})(y</em>{ij} - \bar{y}_i - \bar{y}<em>j + \bar{y}</em>{..})^\prime$</td>
</tr>
<tr>
<td>Total</td>
<td>$vr-1$</td>
<td>$T = \sum_{i=1}^{v} \sum_{j=1}^{b} (y_{ij} - \bar{y}<em>{..})(y</em>{ij} - \bar{y}_{..})^\prime = H + B + R$</td>
</tr>
</tbody>
</table>

Here $H$, $B$, $R$ and $T$ are the sum of squares and sum of cross product matrices of treatments, replications, errors (residuals) and totals respectively. The residual sum of squares and cross products matrix for the reduced model $\Omega_0$ is denoted by $R_0$ and is given by

$$R_0 = R + H$$

We reject the null hypothesis of equality of treatment mean vectors if the ratio of generalized variance ($Wilk's$ $lambda$ statistic) $\Lambda = \frac{|R|}{|H + R|}$ is too small. Assuming the normal distribution,
under null hypothesis $A$ is distributed as the product of independent beta variables. A better but more complicated approximation of the distribution of $A$ is

$$\frac{1 - A^{1/b}}{A^{1/b}} \frac{(ab - c)}{ph} \sim F(ph, ab-c)$$

where $a = \left( s - \frac{p - h + 1}{2} \right)$, $b = \sqrt{\left( \frac{p^2 h^2 - 4}{p^2 + h^2 - 5} \right)}$, $c = \frac{ph^2 - 2}{2}$

For some particular values of $h$ and $p$, it reduces to exact F-distribution. The special cases are given below:

For $h = 1$ and any $p$, this reduces to

$$\frac{(1 - \Lambda)(s - p + 1)}{\Lambda} \sim F(p, s - p + 1)$$

For $h = 2$ and any $p$, it reduces to

$$\frac{(1 - \sqrt{\Lambda})(s - p + 1)}{\sqrt{\Lambda}} \sim F(2p, 2(s - p + 1))$$

For $p = 2$ and any $h$

$$\frac{(1 - \sqrt{\Lambda})(s - 1)}{\sqrt{\Lambda}} \sim F(2h, 2(s - 1))$$

For $p = 1$, the statistic reduces to the usual variance ratio statistics.

*The hypothesis regarding the equality of replication effects can be tested by replacing $\Lambda$ by $\frac{|R|}{|B + R|}$ and $h$ by $t$ in the above.*

*Remark:* One complication of multivariate analysis that does not arise in the univariate case is due to the ranks of the matrices. The rank of $R$ should not be smaller than $p$ or in other words error degrees of freedom $s$ should be greater than or equal to $p$ ($s \geq p$).
4. Principal Component Analysis

The purpose of principal component analysis is to derive a small number of linear combinations (principal components) of a set of variables that retain as much information in the original variables as possible. Often a small number of principal components can be used in place of the original variables for plotting, regression, clustering and so on. Principal component analysis can also be viewed as a technique to remove multicollinearity in the data. In this technique, we transform the original set of variables to a new set of uncorrelated random variables. These new variables are linear combinations of the originals variables and are derived in decreasing order of importance so that the first principal component accounts for as much as possible of the variation in the original data.

Let $x_1, x_2, x_3, ..., x_p$ are variables under study, then first principal component may be defined as

$$z_1 = a_{11}x_1 + a_{12}x_2 + ... + a_{1p}x_p$$

such that variance of $z_1$ is as large as possible subject to the condition that

$$a_{11}^2 + a_{12}^2 + ... + a_{1p}^2 = 1$$

This constraint is introduced because if this is not done, then $Var(z_1)$ can be increased simply by multiplying any $a_{1j}$'s by a constant factor. The second principal component is defined as

$$z_2 = a_{21}x_1 + a_{22}x_2 + ... + a_{2p}x_p$$

such that $Var(z_2)$ is as large as possible next to $Var(z_1)$ subject to the constraint that

$$a_{21}^2 + a_{22}^2 + ... + a_{2p}^2 = 1 \text{ and } Cov(z_1, z_2) = 0 \text{ and so on.}$$

It is quite likely that first few principal components account for most of the variability in the original data. If so, these few principal components can then replace the initial $p$ variables in subsequent analysis, thus reducing the effective dimensionality of the problem. An analysis of principal components often reveals relationships that were not previously suspected and thereby allows interpretation that would not ordinarily result. However, Principal Components Analysis is more of a mean to an end rather than end in itself because this frequently serves as intermediate steps in much larger investigations by reducing the dimensionality of the problem and providing easier interpretation. It is a mathematical technique, which does not require user to specify the statistical model or assumption about distribution of original variates. It may also be mentioned that principal components are artificial variables and often it is not possible to assign physical meaning to them. Further, since Principal Components Analysis transforms original set of variables to new set of uncorrelated variables. It is worth stressing that if the original variables are uncorrelated, then there is no point in carrying out the Principal Components Analysis. It is important to note here that the principal components depend on the scale of measurement. Conventional way of getting rid of this problem is to use the standardized variables with unit variances.
Example 4: Let us consider the following data on average minimum temperature \( (x_1) \), average relative humidity at 8 hrs. \( (x_2) \), average relative humidity at 14 hrs. \( (x_3) \) and total rainfall in cm. \( (x_4) \) pertaining to Raipur district from 1970 to 1986 for kharif season from 21st May to 7th Oct.

<table>
<thead>
<tr>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( x_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>25.0</td>
<td>86</td>
<td>66</td>
<td>186.49</td>
</tr>
<tr>
<td>24.9</td>
<td>84</td>
<td>66</td>
<td>124.34</td>
</tr>
<tr>
<td>25.4</td>
<td>77</td>
<td>55</td>
<td>98.79</td>
</tr>
<tr>
<td>24.4</td>
<td>82</td>
<td>62</td>
<td>118.88</td>
</tr>
<tr>
<td>22.9</td>
<td>79</td>
<td>53</td>
<td>71.88</td>
</tr>
<tr>
<td>7.7</td>
<td>86</td>
<td>60</td>
<td>111.96</td>
</tr>
<tr>
<td>25.1</td>
<td>82</td>
<td>58</td>
<td>99.74</td>
</tr>
<tr>
<td>24.9</td>
<td>83</td>
<td>63</td>
<td>115.20</td>
</tr>
<tr>
<td>24.9</td>
<td>82</td>
<td>63</td>
<td>100.16</td>
</tr>
<tr>
<td>24.9</td>
<td>78</td>
<td>56</td>
<td>62.38</td>
</tr>
<tr>
<td>24.3</td>
<td>85</td>
<td>67</td>
<td>154.40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( x_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>24.6</td>
<td>79</td>
<td>61</td>
<td>112.71</td>
</tr>
<tr>
<td>24.3</td>
<td>81</td>
<td>58</td>
<td>79.63</td>
</tr>
<tr>
<td>24.6</td>
<td>81</td>
<td>61</td>
<td>125.59</td>
</tr>
<tr>
<td>24.1</td>
<td>85</td>
<td>64</td>
<td>99.87</td>
</tr>
<tr>
<td>24.5</td>
<td>84</td>
<td>63</td>
<td>143.56</td>
</tr>
<tr>
<td>24.0</td>
<td>81</td>
<td>61</td>
<td>114.97</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td></td>
<td>112.97</td>
</tr>
<tr>
<td>S.D.</td>
<td>4.13</td>
<td>2.75</td>
<td>3.97</td>
</tr>
</tbody>
</table>

with the variance co-variance matrix

\[ \Sigma = \begin{bmatrix} 17.02 & -4.12 & 1.54 & 5.14 \\ -4.12 & 7.56 & 8.50 & 54.82 \\ 1.54 & 8.50 & 15.75 & 92.95 \\ 5.14 & 54.82 & 92.95 & 903.87 \end{bmatrix} \]

Find the eigenvalues and eigenvectors of the above matrix. Arrange the eigenvalues in decreasing order. Let the eigenvalues in decreasing order and corresponding eigenvectors are

\[ \lambda_1 = 916.902 \hspace{1cm} a_1 = (0.006, 0.061, 0.103, 0.993) \]
\[ \lambda_2 = 18.375 \hspace{1cm} a_2 = (0.955, -0.296, 0.011, 0.012) \]
\[ \lambda_3 = 7.87 \hspace{1cm} a_3 = (0.141, 0.485, 0.855, -0.119) \]
\[ \lambda_4 = 1.056 \hspace{1cm} a_4 = (0.260, 0.820, -0.509, 0.001) \]
The principal components for this data will be
\[
\begin{align*}
    z_1 &= 0.006x_1 + 0.061x_2 + 0.103x_3 + 0.993x_4 \\
    z_2 &= 0.955x_1 - 0.296x_2 + 0.011x_3 + 0.012x_4 \\
    z_3 &= 0.141x_1 + 0.485x_2 + 0.855x_3 - 0.119x_4 \\
    z_4 &= 0.26x_1 + 0.82x_2 - 0.509x_3 + 0.001x_4
\end{align*}
\]

The variance of principal components will be eigenvalues i.e.
\[
\begin{align*}
    \text{Var}(z_1) = 916.902, \quad \text{Var}(z_2) = 18.375, \quad \text{Var}(z_3) = 7.87, \quad \text{Var}(z_4) = 1.056
\end{align*}
\]

The total variation explained by principal components is
\[
\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 916.902 + 18.375 + 7.87 + 1.056 = 944.20
\]

As such, it can be seen that the total variation explained by principal components is same as that explained by original variables. It could also be proved mathematically as well as empirically that the principal components are uncorrelated.

The proportion of total variation accounted for by the principal components is
\[
\frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4} = \frac{916.902}{944.203} = 0.97
\]

Continuing, the first two components account for a proportion
\[
\frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4} = \frac{935.277}{944.203} = 0.99
\]

of the total variance.

Hence, in further analysis, the first or first two principal components \( z_1 \) and \( z_2 \) could replace four variables by sacrificing negligible information about the total variation in the system. The scores of principal components can be obtained by substituting the values of \( x_i \)'s in the equations of \( z_i \)'s. For above data, the first two principal components for first observation i.e. for year 1970 can be worked out as
\[
\begin{align*}
    z_1 &= 0.006 \times 25.0 + 0.061 \times 86 + 0.103 \times 66 + 0.993 \times 186.49 - 197.380 \\
    z_2 &= 0.955 \times 25.0 - 0.296 \times 86 + 0.011 \times 66 + 0.012 \times 186.49 - 1.383
\end{align*}
\]

Similarly for the year 1971
\[
\begin{align*}
    z_1 &= 0.006 \times 24.9 + 0.061 \times 84 + 0.103 \times 66 + 0.993 \times 124.34 - 135.54 \\
    z_2 &= 0.955 \times 24.9 - 0.296 \times 84 + 0.011 \times 66 + 0.012 \times 124.34 - 1.134
\end{align*}
\]

Thus the whole data with four variables can be converted to a new data set with two principal components.
Example 5: Consider the same data as given in Example 1. The variance-covariance matrix was given as

$$
\Sigma = \begin{bmatrix}
2.879368 & 10.01 & -1.80905 \\
10.01 & 199.7884 & -5.64 \\
-1.80905 & -5.64 & 3.627658
\end{bmatrix}
$$

Now find the eigenvalues and eigenvectors of the above matrix. Arrange the eigenvalues in decreasing order. Let the eigenvalues in decreasing order and corresponding eigenvectors are

$$
\lambda_1 = 200.462 \quad a_1 = (0.0508, 0.9983, -0.0291) \\
\lambda_2 = 4.532 \quad a_2 = (-0.5737, 0.0530, 0.8173) \\
\lambda_3 = 1.301 \quad a_3 = (0.8175, -0.0249, 0.5754)
$$

The principal components for this data are

$$
z_1 = 0.0508x_1 + 0.9983x_2 - 0.0291x_3 \\
z_2 = -0.5737x_1 + 0.0530x_2 + 0.8173x_3 \\
z_3 = 0.8175x_1 - 0.0249x_2 + 0.5754x_3
$$

The variance of principal components will be eigenvalues i.e.

$$
Var(z_1) = 200.462, \ Var(z_2) = 4.532, \ Var(z_3) = 1.301
$$

The total variation explained by principal components is

$$
\lambda_1 + \lambda_2 + \lambda_3 = 200.462 + 4.532 + 1.301 = 206.295
$$

As such, it can be seen that the total variation explained by principal components is same as that explained by original variables. It could also be proved mathematically as well as empirically that the principal components are uncorrelated.

The proportion of total variation accounted for by the principal components is

$$
\frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} = \frac{200.462}{206.295} = 0.9717 \text{ of the total variance.}
$$

Continuing, the first two components account for a proportion

$$
\frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3} = \frac{204.994}{206.295} = 0.9937 \text{ of the total variance.}
$$

Hence, in further analysis, the first or first two principal components $z_1$ and $z_2$ could replace four variables by sacrificing negligible information about the total variation in the system. The scores of principal components can be obtained by substituting the values of $x_i$'s in the equations of $z_i$'s. For above data, the first two principal components for first observation i.e. for first individual is

$$
z_1 = 0.0508 \times 3.7 + 0.9983 \times 48.5 - 0.0291 \times 9.3 \\
z_2 = -0.5737 \times 3.7 + 0.0530 \times 48.5 + 0.8173 \times 9.3
$$
Similarly principal component scores for other individuals can be obtained. Thus the whole data with three variables can be converted to a new data set with two principal components.

5. **Canonical Correlation Analysis**

Canonical correlation is a technique for analyzing the relationship between two sets of variables. Each set can contain several variables. Simple and multiple correlation are special cases of canonical correlation in which one or both sets contain a single variable. This analysis actually focuses on the correlation between a linear combination of the variables in one set and a linear combination of the variables in the second set. The idea is first to determine the pair of linear combinations having the largest correlation. Next we determine the pair of linear combinations having the largest correlation among all pairs uncorrelated with the initially selected pair. This process continues until the number of pairs of canonical variables equals the number of variables in the smaller group. The pairs of linear combinations are called the **canonical variables** and their correlations are called **canonical correlations**. The canonical correlations measure the strength of association between the two sets of variables. The maximization aspect of the technique represents an attempt to concentrate a high-dimensional relationship between two sets of variables into a few pair of canonical variables.

6. **Discriminant Analysis**

The term discriminant analysis refers to several types of analysis viz. classificatory discriminant analysis (used to classify observations into two or more known groups on the basis of one or more quantitative variables), Canonical discriminant analysis (a dimension reduction technique related to principal components and canonical correlation), Stepwise discriminant analysis (a variable selection technique i.e. to try to find a subset of quantitative variables that best reveals differences among the classes).

For classificatory discriminant analysis, Fisher's Discriminant function is generally used. It is described in the sequel.

Fisher's idea was to transform the multivariate observations $x$ to univariate observations $y$ such the $y$'s derived from the populations $\pi_1$ and $\pi_2$ were separated as much as possible. Fisher's approach assumes that the populations are normal and also assumes the population covariance matrices are equal because a pooled estimate of common covariance matrix is used.
A fixed linear combination of the x's takes the values \( y_1, y_2, \ldots, y_{1n} \) for the observations from the first population and the values \( y_2, y_2, \ldots, y_{2n_2} \) for the observations from the second population. The separation of these two sets of univariate y's is assessed in terms of the differences between \( \bar{y}_1 \) and \( \bar{y}_2 \) expressed in standard deviation units. That is,

\[
\text{separation} = \left( \frac{\bar{y}_1 - \bar{y}_2}{s_y} \right) = \frac{\sum_{j=1}^{n_1} (y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n_2} (y_{2j} - \bar{y}_2)^2}{n_1 + n_2 - 2},
\]

is the pooled estimate of the variance. The objective is to select the linear combination of the x to achieve maximum separation of the sample means \( \bar{y}_1 \) and \( \bar{y}_2 \).

Result: The linear combination \( y = \hat{y}x = (\bar{x}_1 - \bar{x}_2)' S_{\text{pooled}}^{-1} x \) maximizes the ratio

\[
\frac{\text{(Squared distance between sample mean of y)}}{\text{(Sample variance of y)}} = \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} = \frac{(\hat{y}' d)^2}{\hat{y}' S_{\text{pooled}} \hat{y}}
\]

over all possible coefficient vectors \( \hat{y} \) where \( d = (\bar{x}_1 - \bar{x}_2) \). The maximum of the above ratio is \( D^2 = (\bar{x}_1 - \bar{x}_2)' S_{\text{pooled}}^{-1} (\bar{x}_1 - \bar{x}_2) \), the Mahalanobis distance.

Fisher's solution to the separation problem can also be used to classify new observations. An allocation rule is as follows.

Allocate \( x_0 \) to \( \pi_1 \) if \( y_0 = (\bar{x}_1 - \bar{x}_2)' S_{\text{pooled}}^{-1} x_0 \geq \hat{m} = \frac{1}{2} (\bar{x}_1 - \bar{x}_2)' S_{\text{pooled}}^{-1} (\bar{x}_1 + \bar{x}_2) \)

and to \( \pi_2 \) if \( y_0 < \hat{m} \)

If we assume the populations \( \pi_1 \) and \( \pi_2 \) are multivariate normal with a common covariance matrix, the a test of \( H_0: \mu_1 = \mu_2 \) versus \( H_1: \mu_1 \neq \mu_2 \) is accomplished by referring

\[
\left( \frac{n_1 + n_2 - p - 1}{n_1 + n_2 - 2} \right) \left( \frac{n_1 n_2}{n_1 + n_2} \right) \chi^2
\]
to an F-distribution with $v_1 = p$ and $v_2 = n_1 + n_2 - p - 1$ degrees of freedom. If $H_0$ is rejected, we can conclude the separation between the two populations is significant.

**Example 6:** (Example 11.3 in Johnson and Wichern, 2002). To construct a procedure for detecting potential hemophilia 'A' carriers, blood samples were analyzed for two groups of women and measurements on two variables, $x_1 = \log_{10}(AHF \text{ activity})$ and $x_2 = \log_{10}(AHF-like \text{ antigens})$ recorded. The first group of $n_1 = 30$ women were selected from a population who do not carry hemophilia gene (normal group). The second group of $n_2 = 22$ women were selected from known hemophilia 'A' carriers (obligatory group). The mean vectors and sample covariance matrix are given as

$$
\bar{x}_1 = \begin{bmatrix} -0.0065 \\ -0.0390 \end{bmatrix}, \quad \bar{x}_2 = \begin{bmatrix} -0.2483 \\ 0.0262 \end{bmatrix} \quad \text{and} \quad S^{-1}_{\text{pooled}} = \begin{bmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{bmatrix}
$$

Now the linear discriminant function is

$$y_0 = \hat{x}_0 = (\bar{x}_1 - \bar{x}_2)' S^{-1}_{\text{pooled}} x_0$$

$$= \begin{bmatrix} 0.2418 & -0.0652 \end{bmatrix} \begin{bmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= 37.61 x_1 - 28.92 x_2$$

Moreover

$$\bar{y}_1 = \hat{x}_1 = \begin{bmatrix} 37.61 \\ -28.92 \end{bmatrix} \begin{bmatrix} -0.0065 \\ -0.0390 \end{bmatrix} = 0.88$$

$$\bar{y}_2 = \hat{x}_2 = \begin{bmatrix} 37.61 \\ -28.92 \end{bmatrix} \begin{bmatrix} -0.2483 \\ 0.0262 \end{bmatrix} = -10.10$$

And the mid-point between these means is

$$\hat{m} = \frac{1}{2} (\bar{x}_1 - \bar{x}_2) S^{-1}_{\text{pooled}} (\bar{x}_1 + \bar{x}_2) = \frac{1}{2} (\bar{y}_1 + \bar{y}_2) = -4.61$$

Now to classify a women who may be a hemophilia 'A' carrier with $x_1 = -0.210$ and $x_2 = -0.044$.

We calculate: $y_0 = \hat{x}_0 = 37.61 x_1 - 28.92 x_2 = -6.62$. Since $y_0 < \hat{m}$ we classify the women in $\pi_2$ population, i.e., to obligatory carrier group.
7. Factor Analysis

The essential purpose of factor analysis is to describe, if possible, the covariance relationships among many variables in terms of a few underlying but unobservable random quantities called factors. A frequent source of confusion in the field of factor analysis is the term factor. It sometimes refers to a hypothetical, unobservable variable as in the phrase common factor. In this sense, factor analysis must be distinguished from component analysis since a component is an observable linear combination. Factor is also used in the sense of matrix factor, in that one matrix is a factor of second matrix if the first matrix multiplied by its transpose equals the second matrix. In this sense, factor analysis refers to all methods of data analysis using matrix factors, including component analysis and common factor analysis. A common factor is an unobservable hypothetical variable that contributes to that variance of at least two of the observed variables. The unqualified term “factor” often refers to a common factor. A unique factor is an unobservable hypothetical variable that contributes to the variance of only one of the observed variables. The model for common factor analysis posits one unique factor for each observed variable.

Example 7: What underlying attitudes lead people to respond to the questions on a political survey as they do? Examining the correlations among the survey items reveals that there is significant overlap among various subgroups of items—questions about taxes tend to correlate with each other, questions about military issues correlate with each other, and so on. With factor analysis, you can investigate the number of underlying factors and, in many cases, you can identify what the factors represent conceptually. Additionally, you can compute factor scores for each respondent, which can then be used in subsequent analyses. For example, you might build a logistic regression model to predict voting behavior based on factor scores.

Example 8: A manufacturer of fabricating parts is interested in identifying the determinants of a successful salesperson. The manufacturer has on file the information shown in the following table. He is wondering whether he could reduce these seven variables to two or three factors, for a meaningful appreciation of the problem.
### Data Matrix for Factor Analysis of seven variables (14 sales people)

<table>
<thead>
<tr>
<th>Sales Person</th>
<th>Height ((x_1))</th>
<th>Weight ((x_2))</th>
<th>Education ((x_3))</th>
<th>Age ((x_4))</th>
<th>No. of Children ((x_5))</th>
<th>Size of Household ((x_6))</th>
<th>IQ ((x_7))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>67</td>
<td>155</td>
<td>12</td>
<td>27</td>
<td>0</td>
<td>2</td>
<td>102</td>
</tr>
<tr>
<td>2</td>
<td>69</td>
<td>175</td>
<td>11</td>
<td>35</td>
<td>3</td>
<td>6</td>
<td>92</td>
</tr>
<tr>
<td>3</td>
<td>71</td>
<td>170</td>
<td>14</td>
<td>32</td>
<td>1</td>
<td>3</td>
<td>111</td>
</tr>
<tr>
<td>4</td>
<td>70</td>
<td>160</td>
<td>16</td>
<td>25</td>
<td>0</td>
<td>1</td>
<td>115</td>
</tr>
<tr>
<td>5</td>
<td>72</td>
<td>180</td>
<td>12</td>
<td>36</td>
<td>2</td>
<td>4</td>
<td>108</td>
</tr>
<tr>
<td>6</td>
<td>69</td>
<td>170</td>
<td>11</td>
<td>41</td>
<td>3</td>
<td>5</td>
<td>90</td>
</tr>
<tr>
<td>7</td>
<td>74</td>
<td>195</td>
<td>13</td>
<td>36</td>
<td>1</td>
<td>2</td>
<td>114</td>
</tr>
<tr>
<td>8</td>
<td>68</td>
<td>160</td>
<td>16</td>
<td>32</td>
<td>1</td>
<td>3</td>
<td>118</td>
</tr>
<tr>
<td>9</td>
<td>70</td>
<td>175</td>
<td>12</td>
<td>45</td>
<td>4</td>
<td>6</td>
<td>121</td>
</tr>
<tr>
<td>10</td>
<td>71</td>
<td>180</td>
<td>13</td>
<td>24</td>
<td>0</td>
<td>2</td>
<td>92</td>
</tr>
<tr>
<td>11</td>
<td>66</td>
<td>145</td>
<td>10</td>
<td>39</td>
<td>2</td>
<td>4</td>
<td>100</td>
</tr>
<tr>
<td>12</td>
<td>75</td>
<td>210</td>
<td>16</td>
<td>26</td>
<td>0</td>
<td>1</td>
<td>109</td>
</tr>
<tr>
<td>13</td>
<td>70</td>
<td>160</td>
<td>12</td>
<td>31</td>
<td>0</td>
<td>3</td>
<td>102</td>
</tr>
<tr>
<td>14</td>
<td>71</td>
<td>175</td>
<td>13</td>
<td>43</td>
<td>3</td>
<td>5</td>
<td>112</td>
</tr>
</tbody>
</table>

Can we now collapse the seven variables into three factors? Intuition might suggest the presence of three primary factors: maturity revealed in age/children/size of household, physical size as shown by height and weight, and intelligence or training as revealed by education and IQ.

### Three-factor results with seven variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Factor I (\text{Factor}_I)</th>
<th>Factor II (\text{Factor}_II)</th>
<th>Factor III (\text{Factor}_III)</th>
<th>Communality (\text{Communality})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>0.59038</td>
<td>0.72170</td>
<td>-0.30331</td>
<td>0.96140 (sumsq I,II and III)</td>
</tr>
<tr>
<td>Weight</td>
<td>0.45256</td>
<td>0.75932</td>
<td>-0.44273</td>
<td>0.97738</td>
</tr>
<tr>
<td>Education</td>
<td>0.80252</td>
<td>0.18513</td>
<td>0.42631</td>
<td>0.86006</td>
</tr>
<tr>
<td>Age</td>
<td>-0.86689</td>
<td>0.41116</td>
<td>0.18733</td>
<td>0.95564</td>
</tr>
<tr>
<td>No. of Children</td>
<td>-0.84930</td>
<td>0.49247</td>
<td>0.05883</td>
<td>0.96730</td>
</tr>
<tr>
<td>Size of Household</td>
<td>-0.92582</td>
<td>0.30007</td>
<td>-0.01953</td>
<td>0.94756</td>
</tr>
<tr>
<td>IQ</td>
<td>0.28761</td>
<td>0.46696</td>
<td>0.80524</td>
<td>0.94918</td>
</tr>
<tr>
<td>Sum of squares</td>
<td>3.61007</td>
<td>1.85136</td>
<td>1.15709</td>
<td></td>
</tr>
<tr>
<td>Variance summarized</td>
<td>0.51572</td>
<td>0.26448</td>
<td>0.16530</td>
<td>Average=0.94550</td>
</tr>
</tbody>
</table>
Factor Loadings
The coefficients in the factor equations are called "factor loadings". They appear above in each factor column, corresponding to each variable. The equations are:

\[ F_1 = 0.59038x_1 + 0.45256x_2 + 0.80252x_3 - 0.86689x_4 - 0.84930x_5 - 0.92582x_6 + 0.28761x_7 \]
\[ F_2 = 0.72170x_1 + 0.75932x_2 + 0.18513x_3 + 0.41116x_4 + 0.49247x_5 + 0.30007x_6 + 0.46696x_7 \]
\[ F_3 = -0.30331x_1 - 0.44273x_2 + 0.80252x_3 + 0.18733x_4 + 0.58830x_5 - 0.01953x_6 + 0.80324x_7 \]

The factor loadings depict the relative importance of each variable with respect to a particular factor. In all the three equations, education \((x_3)\) and IQ \((x_7)\) have got positive loading factors indicating that they are variables of importance in determining the success of sales person.

Variance summarized
Factor analysis employs the criterion of maximum reduction of variance - variance found in the initial set of variables. Each factor contributes to reduction. In our example Factor I accounts for 51.6% of the total variance. Factor II for 26.4% and Factor III for 16.5%. Together the three factors "explain" almost 95% of the variance.

Communality
In the ideal solution the factors derived will explain 100% of the variance in each of the original variables. "Communality" measures the percentage of the variance in the original variables that is captured by the combinations of factors in the solution. Thus communality is computed for each of the original variables. Each variables communality might be thought of as showing the extent to which it is revealed by the system of factors. In our example the communality is over 85% for every variable. Thus the three factors seem to capture the underlying dimensions involved in these variables.
8. Cluster Analysis

The basic aim of the cluster analysis is to find “natural” or “real” groupings, if any, of a set of individuals (or objects or points or units or whatever). This set of individuals may form a complete population or be a sample from a larger population. More formally, cluster analysis aims to allocate a set of individuals to a set of mutually exclusive, exhaustive groups such that individuals within a group are similar to one another while individuals in different groups are dissimilar. This set of groups is called partition or dissection. Cluster analysis can also be used for summarizing the data rather than finding natural or real groupings. Grouping or clustering is distinct from the classification methods in the sense that the classification pertains to a known number of groups, and the operational objective is to assign new observations to one of these groups. Cluster analysis is a more primitive technique in that no assumptions are made concerning the number of groups or the group structure. Grouping is done on the basis of similarities or distances (dissimilarities). Some of these distance criteria are:

Euclidean distance: This is probably the most commonly chosen type of distance. It is the geometric distance in the multidimensional space and is computed as:

\[
 d(x, y) = \left[ \sum_{i=1}^{p} (x_i - y_i)^2 \right]^{1/2} = \sqrt{(x - y)'(x - y)}
\]

where \( x, y \) are the \( p \)-dimensional vectors of observations.

Note that Euclidean (and squared Euclidean) distances are usually computed from raw data, and not from standardized data. This method has certain advantages (e.g., the distance between any two objects is not affected by the addition of new objects to the analysis, which may be outliers). However, the distances can be greatly affected by differences in scale among the dimensions from which the distances are computed. For example, if one of the dimensions denotes a measured length in centimeters, and you then convert it to millimeters (by multiplying the values by 10), the resulting Euclidean or squared Euclidean distances (computed from multiple dimensions) can be greatly affected (i.e., biased by those dimensions which have a larger scale), and consequently, the results of cluster analyses may be very different. Generally, it is good practice to transform the dimensions so they have similar scales.

Squared Euclidean distance: This measure is used in order to place progressively greater weight on objects that are further apart. This distance is square of the Euclidean distance.

Statistical distance: The statistical distance between the two \( p \)-dimensional vectors \( x \) and \( y \) is 
\[
 d(x, y) = \sqrt{(x - y)'s^{-1}(x - y)}, \text{ where } s \text{ is the sample variance-covariance matrix.}
\]
References
### Appendix A: Critical values for t-distribution

<table>
<thead>
<tr>
<th>DF</th>
<th>Probability %</th>
<th>DF</th>
<th>Probability %</th>
<th>DF</th>
<th>Probability %</th>
<th>DF</th>
<th>Probability %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.01</td>
<td>0.05</td>
<td></td>
<td>0.01</td>
<td>0.05</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>1</td>
<td>63.657</td>
<td>12.706</td>
<td>31</td>
<td>2.744</td>
<td>2.040</td>
<td>61</td>
<td>2.658</td>
</tr>
<tr>
<td>2</td>
<td>9.295</td>
<td>4.303</td>
<td>32</td>
<td>2.738</td>
<td>2.037</td>
<td>62</td>
<td>2.657</td>
</tr>
<tr>
<td>3</td>
<td>5.841</td>
<td>3.182</td>
<td>33</td>
<td>2.733</td>
<td>2.035</td>
<td>63</td>
<td>2.656</td>
</tr>
<tr>
<td>4</td>
<td>4.604</td>
<td>2.776</td>
<td>34</td>
<td>2.728</td>
<td>2.033</td>
<td>64</td>
<td>2.654</td>
</tr>
<tr>
<td>5</td>
<td>4.032</td>
<td>2.571</td>
<td>35</td>
<td>2.723</td>
<td>2.030</td>
<td>65</td>
<td>2.653</td>
</tr>
<tr>
<td>6</td>
<td>3.707</td>
<td>2.447</td>
<td>36</td>
<td>2.719</td>
<td>2.028</td>
<td>66</td>
<td>2.652</td>
</tr>
<tr>
<td>7</td>
<td>3.499</td>
<td>2.365</td>
<td>37</td>
<td>2.715</td>
<td>2.028</td>
<td>67</td>
<td>2.651</td>
</tr>
<tr>
<td>8</td>
<td>3.355</td>
<td>2.306</td>
<td>38</td>
<td>2.711</td>
<td>2.024</td>
<td>68</td>
<td>2.650</td>
</tr>
<tr>
<td>9</td>
<td>3.250</td>
<td>2.262</td>
<td>39</td>
<td>2.707</td>
<td>2.022</td>
<td>69</td>
<td>2.649</td>
</tr>
<tr>
<td>10</td>
<td>3.169</td>
<td>2.228</td>
<td>40</td>
<td>2.704</td>
<td>2.021</td>
<td>70</td>
<td>2.647</td>
</tr>
<tr>
<td>11</td>
<td>3.106</td>
<td>2.201</td>
<td>41</td>
<td>2.701</td>
<td>2.020</td>
<td>71</td>
<td>2.646</td>
</tr>
<tr>
<td>12</td>
<td>3.055</td>
<td>2.179</td>
<td>42</td>
<td>2.698</td>
<td>2.018</td>
<td>72</td>
<td>2.645</td>
</tr>
<tr>
<td>13</td>
<td>3.012</td>
<td>2.160</td>
<td>43</td>
<td>2.695</td>
<td>2.017</td>
<td>73</td>
<td>2.644</td>
</tr>
<tr>
<td>14</td>
<td>2.977</td>
<td>2.145</td>
<td>44</td>
<td>2.692</td>
<td>2.016</td>
<td>74</td>
<td>2.643</td>
</tr>
<tr>
<td>15</td>
<td>2.947</td>
<td>2.131</td>
<td>45</td>
<td>2.689</td>
<td>2.014</td>
<td>75</td>
<td>2.643</td>
</tr>
<tr>
<td>16</td>
<td>2.921</td>
<td>2.120</td>
<td>46</td>
<td>2.687</td>
<td>2.013</td>
<td>76</td>
<td>2.642</td>
</tr>
<tr>
<td>17</td>
<td>2.898</td>
<td>2.110</td>
<td>47</td>
<td>2.684</td>
<td>2.012</td>
<td>77</td>
<td>2.641</td>
</tr>
<tr>
<td>18</td>
<td>2.876</td>
<td>2.101</td>
<td>48</td>
<td>2.682</td>
<td>2.011</td>
<td>78</td>
<td>2.640</td>
</tr>
<tr>
<td>19</td>
<td>2.861</td>
<td>2.093</td>
<td>49</td>
<td>2.679</td>
<td>2.010</td>
<td>79</td>
<td>2.639</td>
</tr>
<tr>
<td>20</td>
<td>2.845</td>
<td>2.086</td>
<td>50</td>
<td>2.677</td>
<td>2.008</td>
<td>80</td>
<td>2.638</td>
</tr>
<tr>
<td>21</td>
<td>2.831</td>
<td>2.080</td>
<td>51</td>
<td>2.675</td>
<td>2.007</td>
<td>81</td>
<td>2.637</td>
</tr>
<tr>
<td>22</td>
<td>2.819</td>
<td>2.074</td>
<td>52</td>
<td>2.673</td>
<td>2.006</td>
<td>82</td>
<td>2.637</td>
</tr>
<tr>
<td>23</td>
<td>2.807</td>
<td>2.069</td>
<td>53</td>
<td>2.671</td>
<td>2.005</td>
<td>83</td>
<td>2.636</td>
</tr>
<tr>
<td>24</td>
<td>2.797</td>
<td>2.064</td>
<td>54</td>
<td>2.670</td>
<td>2.004</td>
<td>84</td>
<td>2.635</td>
</tr>
<tr>
<td>25</td>
<td>2.787</td>
<td>2.060</td>
<td>55</td>
<td>2.668</td>
<td>2.004</td>
<td>85</td>
<td>2.634</td>
</tr>
<tr>
<td>26</td>
<td>2.779</td>
<td>2.056</td>
<td>56</td>
<td>2.666</td>
<td>2.003</td>
<td>86</td>
<td>2.634</td>
</tr>
<tr>
<td>27</td>
<td>2.771</td>
<td>2.052</td>
<td>57</td>
<td>2.664</td>
<td>2.002</td>
<td>87</td>
<td>2.633</td>
</tr>
<tr>
<td>28</td>
<td>2.763</td>
<td>2.048</td>
<td>58</td>
<td>2.663</td>
<td>2.002</td>
<td>88</td>
<td>2.632</td>
</tr>
<tr>
<td>29</td>
<td>2.756</td>
<td>2.045</td>
<td>59</td>
<td>2.661</td>
<td>2.001</td>
<td>89</td>
<td>2.632</td>
</tr>
<tr>
<td>30</td>
<td>2.750</td>
<td>2.042</td>
<td>60</td>
<td>2.660</td>
<td>2.000</td>
<td>90</td>
<td>2.631</td>
</tr>
</tbody>
</table>
Appendix B: Critical values for F-distribution

<table>
<thead>
<tr>
<th>Smaller MS (n₂)</th>
<th>Degrees of freedom for greater mean square (n₁)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>%</td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>161.00</td>
</tr>
<tr>
<td>1%</td>
<td>4052.00</td>
</tr>
<tr>
<td>3</td>
<td>10.13</td>
</tr>
<tr>
<td>1%</td>
<td>34.12</td>
</tr>
<tr>
<td>4</td>
<td>7.71</td>
</tr>
<tr>
<td>1%</td>
<td>21.20</td>
</tr>
<tr>
<td>5</td>
<td>6.61</td>
</tr>
<tr>
<td>1%</td>
<td>16.26</td>
</tr>
<tr>
<td>6</td>
<td>5.99</td>
</tr>
<tr>
<td>1%</td>
<td>13.74</td>
</tr>
<tr>
<td>7</td>
<td>5.59</td>
</tr>
<tr>
<td>1%</td>
<td>12.25</td>
</tr>
<tr>
<td>8</td>
<td>5.32</td>
</tr>
<tr>
<td>1%</td>
<td>11.26</td>
</tr>
<tr>
<td>9</td>
<td>5.12</td>
</tr>
<tr>
<td>1%</td>
<td>10.56</td>
</tr>
<tr>
<td>10</td>
<td>4.96</td>
</tr>
<tr>
<td>1%</td>
<td>10.04</td>
</tr>
<tr>
<td>11</td>
<td>4.84</td>
</tr>
<tr>
<td>1%</td>
<td>9.65</td>
</tr>
<tr>
<td>12</td>
<td>4.75</td>
</tr>
<tr>
<td>1%</td>
<td>9.33</td>
</tr>
<tr>
<td>13</td>
<td>4.67</td>
</tr>
<tr>
<td>1%</td>
<td>9.07</td>
</tr>
<tr>
<td>-----</td>
<td>------</td>
</tr>
<tr>
<td>14%</td>
<td>4.60</td>
</tr>
<tr>
<td>1%</td>
<td>8.86</td>
</tr>
<tr>
<td>15%</td>
<td>4.54</td>
</tr>
<tr>
<td>1%</td>
<td>8.68</td>
</tr>
<tr>
<td>16%</td>
<td>4.49</td>
</tr>
<tr>
<td>1%</td>
<td>8.53</td>
</tr>
<tr>
<td>17%</td>
<td>4.45</td>
</tr>
<tr>
<td>1%</td>
<td>8.40</td>
</tr>
<tr>
<td>18%</td>
<td>4.41</td>
</tr>
<tr>
<td>1%</td>
<td>8.28</td>
</tr>
<tr>
<td>19%</td>
<td>4.38</td>
</tr>
<tr>
<td>1%</td>
<td>8.18</td>
</tr>
<tr>
<td>20%</td>
<td>4.35</td>
</tr>
<tr>
<td>1%</td>
<td>8.10</td>
</tr>
<tr>
<td>21%</td>
<td>4.32</td>
</tr>
<tr>
<td>1%</td>
<td>8.02</td>
</tr>
<tr>
<td>22%</td>
<td>4.30</td>
</tr>
<tr>
<td>1%</td>
<td>7.94</td>
</tr>
<tr>
<td>23%</td>
<td>4.28</td>
</tr>
<tr>
<td>1%</td>
<td>7.88</td>
</tr>
<tr>
<td>24%</td>
<td>4.26</td>
</tr>
<tr>
<td>1%</td>
<td>7.82</td>
</tr>
<tr>
<td>25%</td>
<td>4.24</td>
</tr>
<tr>
<td>1%</td>
<td>7.77</td>
</tr>
<tr>
<td>26%</td>
<td>4.22</td>
</tr>
<tr>
<td>1%</td>
<td>7.72</td>
</tr>
<tr>
<td>27%</td>
<td>4.21</td>
</tr>
</tbody>
</table>
### Appendix B : Critical values for F-distribution (Continued...)

<table>
<thead>
<tr>
<th>Smaller MS (n₁)</th>
<th>Degrees of freedom for greater mean square (n₂)</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>28 5%</td>
<td></td>
<td>4.20</td>
<td>3.34</td>
<td>2.95</td>
<td>2.71</td>
<td>2.56</td>
<td>2.44</td>
<td>2.36</td>
<td>2.29</td>
<td>2.24</td>
<td>2.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.64</td>
<td>5.45</td>
<td>4.57</td>
<td>4.07</td>
<td>3.76</td>
<td>3.53</td>
<td>3.36</td>
<td>3.23</td>
<td>3.11</td>
<td>3.03</td>
</tr>
<tr>
<td>29 5%</td>
<td></td>
<td>4.18</td>
<td>3.33</td>
<td>2.95</td>
<td>2.70</td>
<td>2.54</td>
<td>2.43</td>
<td>2.35</td>
<td>2.28</td>
<td>2.22</td>
<td>2.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.60</td>
<td>5.42</td>
<td>4.54</td>
<td>4.04</td>
<td>3.73</td>
<td>3.50</td>
<td>3.33</td>
<td>3.20</td>
<td>3.08</td>
<td>3.00</td>
</tr>
<tr>
<td>30 5%</td>
<td></td>
<td>4.17</td>
<td>3.32</td>
<td>2.92</td>
<td>2.69</td>
<td>2.53</td>
<td>2.42</td>
<td>2.34</td>
<td>2.27</td>
<td>2.21</td>
<td>2.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.56</td>
<td>5.39</td>
<td>4.51</td>
<td>4.02</td>
<td>3.70</td>
<td>3.47</td>
<td>3.30</td>
<td>3.17</td>
<td>3.06</td>
<td>2.98</td>
</tr>
<tr>
<td>31 5%</td>
<td></td>
<td>4.16</td>
<td>3.31</td>
<td>2.91</td>
<td>2.68</td>
<td>2.52</td>
<td>2.41</td>
<td>2.33</td>
<td>2.26</td>
<td>2.20</td>
<td>2.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.53</td>
<td>5.37</td>
<td>4.49</td>
<td>4.00</td>
<td>3.68</td>
<td>3.45</td>
<td>3.28</td>
<td>3.15</td>
<td>3.04</td>
<td>2.96</td>
</tr>
<tr>
<td>32 5%</td>
<td></td>
<td>4.15</td>
<td>3.30</td>
<td>2.90</td>
<td>2.67</td>
<td>2.51</td>
<td>2.40</td>
<td>2.32</td>
<td>2.25</td>
<td>2.19</td>
<td>2.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.50</td>
<td>5.34</td>
<td>4.46</td>
<td>3.97</td>
<td>3.66</td>
<td>3.42</td>
<td>3.25</td>
<td>3.12</td>
<td>3.01</td>
<td>2.94</td>
</tr>
<tr>
<td>33 5%</td>
<td></td>
<td>4.14</td>
<td>3.29</td>
<td>2.89</td>
<td>2.66</td>
<td>2.50</td>
<td>2.39</td>
<td>2.31</td>
<td>2.24</td>
<td>2.18</td>
<td>2.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.47</td>
<td>5.32</td>
<td>4.44</td>
<td>3.95</td>
<td>3.64</td>
<td>3.40</td>
<td>3.23</td>
<td>3.10</td>
<td>2.99</td>
<td>2.92</td>
</tr>
<tr>
<td>34 5%</td>
<td></td>
<td>4.13</td>
<td>3.28</td>
<td>2.88</td>
<td>2.65</td>
<td>2.49</td>
<td>2.38</td>
<td>2.30</td>
<td>2.23</td>
<td>2.17</td>
<td>2.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.44</td>
<td>5.29</td>
<td>4.42</td>
<td>3.93</td>
<td>3.61</td>
<td>3.38</td>
<td>3.21</td>
<td>3.08</td>
<td>2.97</td>
<td>2.89</td>
</tr>
<tr>
<td>35 5%</td>
<td></td>
<td>4.12</td>
<td>3.27</td>
<td>2.87</td>
<td>2.64</td>
<td>2.49</td>
<td>2.37</td>
<td>2.29</td>
<td>2.22</td>
<td>2.16</td>
<td>2.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.42</td>
<td>5.27</td>
<td>4.40</td>
<td>3.91</td>
<td>3.60</td>
<td>3.37</td>
<td>3.20</td>
<td>3.06</td>
<td>2.96</td>
<td>2.88</td>
</tr>
<tr>
<td>36 5%</td>
<td></td>
<td>4.11</td>
<td>3.26</td>
<td>2.86</td>
<td>2.66</td>
<td>2.48</td>
<td>2.36</td>
<td>2.28</td>
<td>2.21</td>
<td>2.15</td>
<td>2.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.39</td>
<td>5.25</td>
<td>4.38</td>
<td>3.89</td>
<td>3.58</td>
<td>3.35</td>
<td>3.18</td>
<td>3.04</td>
<td>2.94</td>
<td>2.86</td>
</tr>
<tr>
<td>37 5%</td>
<td></td>
<td>4.11</td>
<td>3.26</td>
<td>2.86</td>
<td>2.66</td>
<td>2.47</td>
<td>2.36</td>
<td>2.27</td>
<td>2.20</td>
<td>2.15</td>
<td>2.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.37</td>
<td>5.23</td>
<td>4.36</td>
<td>3.88</td>
<td>3.56</td>
<td>3.34</td>
<td>3.17</td>
<td>3.03</td>
<td>2.93</td>
<td>2.84</td>
</tr>
<tr>
<td>38 5%</td>
<td></td>
<td>4.10</td>
<td>3.25</td>
<td>2.85</td>
<td>2.66</td>
<td>2.46</td>
<td>2.35</td>
<td>2.26</td>
<td>2.19</td>
<td>2.14</td>
<td>2.09</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.35</td>
<td>5.21</td>
<td>4.34</td>
<td>3.86</td>
<td>3.54</td>
<td>3.32</td>
<td>3.15</td>
<td>3.02</td>
<td>2.91</td>
<td>2.82</td>
</tr>
<tr>
<td>39 5%</td>
<td></td>
<td>4.09</td>
<td>3.24</td>
<td>2.85</td>
<td>2.66</td>
<td>2.46</td>
<td>2.35</td>
<td>2.26</td>
<td>2.19</td>
<td>2.13</td>
<td>2.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.33</td>
<td>5.20</td>
<td>4.33</td>
<td>3.85</td>
<td>3.53</td>
<td>3.31</td>
<td>3.14</td>
<td>3.01</td>
<td>2.90</td>
<td>2.81</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<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>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>40 %</td>
<td>4.08</td>
<td>3.23</td>
<td>2.84</td>
<td>2.61</td>
<td>2.45</td>
<td>2.34</td>
<td>2.25</td>
<td>2.18</td>
<td>2.12</td>
<td>2.07</td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>7.31</td>
<td>5.18</td>
<td>4.31</td>
<td>3.83</td>
<td>3.51</td>
<td>3.29</td>
<td>3.12</td>
<td>2.99</td>
<td>2.88</td>
<td>2.80</td>
<td></td>
</tr>
<tr>
<td>41 %</td>
<td>4.08</td>
<td>3.23</td>
<td>2.84</td>
<td>2.61</td>
<td>2.45</td>
<td>2.33</td>
<td>2.25</td>
<td>2.18</td>
<td>2.12</td>
<td>2.07</td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>7.29</td>
<td>5.17</td>
<td>4.30</td>
<td>3.82</td>
<td>3.50</td>
<td>3.28</td>
<td>3.11</td>
<td>2.98</td>
<td>2.87</td>
<td>2.79</td>
<td></td>
</tr>
<tr>
<td>42 %</td>
<td>4.07</td>
<td>3.22</td>
<td>2.83</td>
<td>2.60</td>
<td>2.44</td>
<td>2.32</td>
<td>2.24</td>
<td>2.17</td>
<td>2.11</td>
<td>2.06</td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>7.27</td>
<td>5.15</td>
<td>4.29</td>
<td>3.80</td>
<td>3.49</td>
<td>3.26</td>
<td>3.10</td>
<td>2.96</td>
<td>2.86</td>
<td>2.77</td>
<td></td>
</tr>
<tr>
<td>43 %</td>
<td>4.07</td>
<td>3.22</td>
<td>2.83</td>
<td>2.60</td>
<td>2.44</td>
<td>2.32</td>
<td>2.24</td>
<td>2.17</td>
<td>2.11</td>
<td>2.06</td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>7.26</td>
<td>5.14</td>
<td>4.28</td>
<td>3.79</td>
<td>3.48</td>
<td>3.25</td>
<td>3.09</td>
<td>2.95</td>
<td>2.85</td>
<td>2.76</td>
<td></td>
</tr>
<tr>
<td>44 %</td>
<td>4.06</td>
<td>3.21</td>
<td>2.82</td>
<td>2.59</td>
<td>2.43</td>
<td>2.31</td>
<td>2.23</td>
<td>2.16</td>
<td>2.10</td>
<td>2.05</td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>7.24</td>
<td>5.12</td>
<td>4.26</td>
<td>3.78</td>
<td>3.46</td>
<td>3.24</td>
<td>3.07</td>
<td>2.94</td>
<td>2.84</td>
<td>2.75</td>
<td></td>
</tr>
<tr>
<td>45 %</td>
<td>4.06</td>
<td>3.21</td>
<td>2.82</td>
<td>2.59</td>
<td>2.43</td>
<td>2.31</td>
<td>2.23</td>
<td>2.15</td>
<td>2.10</td>
<td>2.05</td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>7.23</td>
<td>5.11</td>
<td>4.25</td>
<td>3.77</td>
<td>3.45</td>
<td>3.23</td>
<td>3.06</td>
<td>2.93</td>
<td>2.83</td>
<td>2.74</td>
<td></td>
</tr>
<tr>
<td>46 %</td>
<td>4.05</td>
<td>3.20</td>
<td>2.81</td>
<td>2.58</td>
<td>2.42</td>
<td>2.30</td>
<td>2.22</td>
<td>2.14</td>
<td>2.09</td>
<td>2.04</td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>7.21</td>
<td>5.10</td>
<td>4.24</td>
<td>3.76</td>
<td>3.44</td>
<td>3.22</td>
<td>3.05</td>
<td>2.92</td>
<td>2.82</td>
<td>2.73</td>
<td></td>
</tr>
<tr>
<td>47 %</td>
<td>4.05</td>
<td>3.20</td>
<td>2.81</td>
<td>2.58</td>
<td>2.42</td>
<td>2.30</td>
<td>2.22</td>
<td>2.14</td>
<td>2.09</td>
<td>2.04</td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>7.20</td>
<td>5.09</td>
<td>4.23</td>
<td>3.75</td>
<td>3.43</td>
<td>3.21</td>
<td>3.05</td>
<td>2.91</td>
<td>2.81</td>
<td>2.72</td>
<td></td>
</tr>
<tr>
<td>48 %</td>
<td>4.04</td>
<td>3.19</td>
<td>2.80</td>
<td>2.57</td>
<td>2.41</td>
<td>2.30</td>
<td>2.21</td>
<td>2.14</td>
<td>2.08</td>
<td>2.03</td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>7.19</td>
<td>5.08</td>
<td>4.22</td>
<td>3.74</td>
<td>3.42</td>
<td>3.20</td>
<td>3.04</td>
<td>2.90</td>
<td>2.80</td>
<td>2.71</td>
<td></td>
</tr>
<tr>
<td>49 %</td>
<td>4.04</td>
<td>3.19</td>
<td>2.80</td>
<td>2.57</td>
<td>2.41</td>
<td>2.30</td>
<td>2.21</td>
<td>2.14</td>
<td>2.08</td>
<td>2.03</td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>7.18</td>
<td>5.07</td>
<td>4.21</td>
<td>3.73</td>
<td>3.42</td>
<td>3.19</td>
<td>3.03</td>
<td>2.89</td>
<td>2.79</td>
<td>2.71</td>
<td></td>
</tr>
<tr>
<td>50 %</td>
<td>4.03</td>
<td>3.18</td>
<td>2.79</td>
<td>2.56</td>
<td>2.40</td>
<td>2.29</td>
<td>2.20</td>
<td>2.13</td>
<td>2.07</td>
<td>2.02</td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>7.17</td>
<td>5.06</td>
<td>4.20</td>
<td>3.72</td>
<td>3.41</td>
<td>3.18</td>
<td>3.02</td>
<td>2.88</td>
<td>2.78</td>
<td>2.70</td>
<td></td>
</tr>
</tbody>
</table>
### Appendix B

**Critical values for F-distribution (Continued...)**

<table>
<thead>
<tr>
<th>Smaller MS (n₁)</th>
<th>Degrees of freedom for greater mean square (n₂)</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>55 5%</td>
<td></td>
<td>4.02</td>
<td>3.17</td>
<td>2.78</td>
<td>2.54</td>
<td>2.38</td>
<td>2.27</td>
<td>2.18</td>
<td>2.11</td>
<td>2.05</td>
<td>2.00</td>
</tr>
<tr>
<td></td>
<td>1%</td>
<td>7.12</td>
<td>5.01</td>
<td>4.16</td>
<td>3.68</td>
<td>3.37</td>
<td>3.15</td>
<td>2.98</td>
<td>2.85</td>
<td>2.75</td>
<td>2.66</td>
</tr>
<tr>
<td>60 5%</td>
<td></td>
<td>4.00</td>
<td>3.15</td>
<td>2.76</td>
<td>2.52</td>
<td>2.37</td>
<td>2.25</td>
<td>2.17</td>
<td>2.10</td>
<td>2.04</td>
<td>1.99</td>
</tr>
<tr>
<td></td>
<td>1%</td>
<td>7.08</td>
<td>4.98</td>
<td>4.13</td>
<td>3.65</td>
<td>3.34</td>
<td>3.12</td>
<td>2.95</td>
<td>2.82</td>
<td>2.72</td>
<td>2.63</td>
</tr>
<tr>
<td>65 5%</td>
<td></td>
<td>3.99</td>
<td>3.14</td>
<td>2.75</td>
<td>2.51</td>
<td>2.36</td>
<td>2.24</td>
<td>2.15</td>
<td>2.08</td>
<td>2.02</td>
<td>1.98</td>
</tr>
<tr>
<td></td>
<td>1%</td>
<td>7.04</td>
<td>4.95</td>
<td>4.10</td>
<td>3.62</td>
<td>3.31</td>
<td>3.09</td>
<td>2.93</td>
<td>2.79</td>
<td>2.70</td>
<td>2.61</td>
</tr>
<tr>
<td>70 5%</td>
<td></td>
<td>3.98</td>
<td>3.13</td>
<td>2.74</td>
<td>2.50</td>
<td>2.35</td>
<td>2.23</td>
<td>2.14</td>
<td>2.07</td>
<td>2.01</td>
<td>1.97</td>
</tr>
<tr>
<td></td>
<td>1%</td>
<td>7.01</td>
<td>4.92</td>
<td>4.08</td>
<td>3.60</td>
<td>3.29</td>
<td>3.07</td>
<td>2.91</td>
<td>2.77</td>
<td>2.67</td>
<td>2.57</td>
</tr>
<tr>
<td>80 5%</td>
<td></td>
<td>3.96</td>
<td>3.11</td>
<td>2.72</td>
<td>2.48</td>
<td>2.33</td>
<td>2.21</td>
<td>2.12</td>
<td>2.05</td>
<td>1.99</td>
<td>1.95</td>
</tr>
<tr>
<td></td>
<td>1%</td>
<td>6.96</td>
<td>4.88</td>
<td>4.04</td>
<td>3.56</td>
<td>3.25</td>
<td>3.04</td>
<td>2.87</td>
<td>2.74</td>
<td>2.65</td>
<td>2.56</td>
</tr>
<tr>
<td>100 5%</td>
<td></td>
<td>3.94</td>
<td>3.09</td>
<td>2.70</td>
<td>2.46</td>
<td>2.30</td>
<td>2.19</td>
<td>2.10</td>
<td>2.03</td>
<td>1.97</td>
<td>1.92</td>
</tr>
<tr>
<td></td>
<td>1%</td>
<td>6.90</td>
<td>4.82</td>
<td>3.98</td>
<td>3.51</td>
<td>3.20</td>
<td>2.99</td>
<td>2.82</td>
<td>2.69</td>
<td>2.59</td>
<td>2.51</td>
</tr>
<tr>
<td>125 5%</td>
<td></td>
<td>3.92</td>
<td>3.07</td>
<td>2.68</td>
<td>2.44</td>
<td>2.29</td>
<td>2.17</td>
<td>2.08</td>
<td>2.01</td>
<td>1.95</td>
<td>1.90</td>
</tr>
<tr>
<td></td>
<td>1%</td>
<td>6.84</td>
<td>4.78</td>
<td>3.94</td>
<td>3.47</td>
<td>3.17</td>
<td>2.95</td>
<td>2.79</td>
<td>2.65</td>
<td>2.56</td>
<td>2.47</td>
</tr>
<tr>
<td>150 5%</td>
<td></td>
<td>3.91</td>
<td>3.06</td>
<td>2.67</td>
<td>2.43</td>
<td>2.27</td>
<td>2.16</td>
<td>2.07</td>
<td>2.00</td>
<td>1.94</td>
<td>1.89</td>
</tr>
<tr>
<td></td>
<td>1%</td>
<td>6.81</td>
<td>4.75</td>
<td>3.91</td>
<td>3.44</td>
<td>3.14</td>
<td>2.92</td>
<td>2.76</td>
<td>2.62</td>
<td>2.53</td>
<td>2.44</td>
</tr>
<tr>
<td>200 5%</td>
<td></td>
<td>3.89</td>
<td>3.04</td>
<td>2.65</td>
<td>2.41</td>
<td>2.26</td>
<td>2.14</td>
<td>2.05</td>
<td>1.98</td>
<td>1.92</td>
<td>1.87</td>
</tr>
<tr>
<td></td>
<td>1%</td>
<td>6.76</td>
<td>4.71</td>
<td>3.88</td>
<td>3.41</td>
<td>3.11</td>
<td>2.90</td>
<td>2.73</td>
<td>2.60</td>
<td>2.50</td>
<td>2.41</td>
</tr>
<tr>
<td>400 5%</td>
<td></td>
<td>3.86</td>
<td>3.02</td>
<td>2.62</td>
<td>2.39</td>
<td>2.23</td>
<td>2.12</td>
<td>2.03</td>
<td>1.96</td>
<td>1.90</td>
<td>1.85</td>
</tr>
<tr>
<td></td>
<td>1%</td>
<td>6.70</td>
<td>4.66</td>
<td>3.83</td>
<td>3.36</td>
<td>3.06</td>
<td>2.85</td>
<td>2.69</td>
<td>2.55</td>
<td>2.46</td>
<td>2.37</td>
</tr>
<tr>
<td>1000∞</td>
<td></td>
<td>3.85</td>
<td>3.00</td>
<td>2.61</td>
<td>2.38</td>
<td>2.22</td>
<td>2.10</td>
<td>2.02</td>
<td>1.95</td>
<td>1.89</td>
<td>1.84</td>
</tr>
<tr>
<td></td>
<td>1%</td>
<td>6.66</td>
<td>4.62</td>
<td>3.80</td>
<td>3.34</td>
<td>3.04</td>
<td>2.82</td>
<td>2.66</td>
<td>2.53</td>
<td>2.43</td>
<td>2.34</td>
</tr>
<tr>
<td>∞ 5%</td>
<td></td>
<td>3.84</td>
<td>2.99</td>
<td>2.60</td>
<td>2.37</td>
<td>2.21</td>
<td>2.09</td>
<td>2.01</td>
<td>1.94</td>
<td>1.88</td>
<td>1.83</td>
</tr>
<tr>
<td></td>
<td>1%</td>
<td>6.64</td>
<td>4.60</td>
<td>3.78</td>
<td>3.32</td>
<td>3.02</td>
<td>2.80</td>
<td>2.64</td>
<td>2.51</td>
<td>2.41</td>
<td>2.32</td>
</tr>
</tbody>
</table>


Appendix C : Critical value for Correlation coefficients (Simple or Partial)

<table>
<thead>
<tr>
<th>DF</th>
<th>Probability %</th>
<th>DF</th>
<th>Probability %</th>
<th>DF</th>
<th>Probability %</th>
<th>DF</th>
<th>Probability %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.01</td>
<td></td>
<td>0.05</td>
<td></td>
<td>0.01</td>
<td></td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>1.000</td>
<td>26</td>
<td>0.478</td>
<td>52</td>
<td>0.348</td>
<td>105</td>
<td>0.248</td>
</tr>
<tr>
<td>2</td>
<td>0.990</td>
<td>27</td>
<td>0.470</td>
<td>54</td>
<td>0.341</td>
<td>110</td>
<td>0.242</td>
</tr>
<tr>
<td>3</td>
<td>0.959</td>
<td>28</td>
<td>0.463</td>
<td>56</td>
<td>0.336</td>
<td>115</td>
<td>0.237</td>
</tr>
<tr>
<td>4</td>
<td>0.917</td>
<td>29</td>
<td>0.456</td>
<td>58</td>
<td>0.330</td>
<td>120</td>
<td>0.232</td>
</tr>
<tr>
<td>5</td>
<td>0.874</td>
<td>30</td>
<td>0.449</td>
<td>60</td>
<td>0.325</td>
<td>125</td>
<td>0.228</td>
</tr>
<tr>
<td>6</td>
<td>0.834</td>
<td>31</td>
<td>0.442</td>
<td>62</td>
<td>0.320</td>
<td>130</td>
<td>0.223</td>
</tr>
<tr>
<td>7</td>
<td>0.798</td>
<td>32</td>
<td>0.436</td>
<td>64</td>
<td>0.315</td>
<td>135</td>
<td>0.219</td>
</tr>
<tr>
<td>8</td>
<td>0.765</td>
<td>33</td>
<td>0.430</td>
<td>66</td>
<td>0.310</td>
<td>140</td>
<td>0.215</td>
</tr>
<tr>
<td>9</td>
<td>0.735</td>
<td>34</td>
<td>0.424</td>
<td>68</td>
<td>0.306</td>
<td>145</td>
<td>0.212</td>
</tr>
<tr>
<td>10</td>
<td>0.708</td>
<td>35</td>
<td>0.418</td>
<td>70</td>
<td>0.302</td>
<td>150</td>
<td>0.208</td>
</tr>
<tr>
<td>11</td>
<td>0.684</td>
<td>36</td>
<td>0.413</td>
<td>72</td>
<td>0.298</td>
<td>160</td>
<td>0.202</td>
</tr>
<tr>
<td>12</td>
<td>0.661</td>
<td>37</td>
<td>0.408</td>
<td>74</td>
<td>0.294</td>
<td>170</td>
<td>0.196</td>
</tr>
<tr>
<td>13</td>
<td>0.641</td>
<td>38</td>
<td>0.403</td>
<td>76</td>
<td>0.290</td>
<td>180</td>
<td>0.190</td>
</tr>
<tr>
<td>14</td>
<td>0.623</td>
<td>39</td>
<td>0.398</td>
<td>78</td>
<td>0.286</td>
<td>190</td>
<td>0.185</td>
</tr>
<tr>
<td>15</td>
<td>0.606</td>
<td>40</td>
<td>0.393</td>
<td>80</td>
<td>0.283</td>
<td>200</td>
<td>0.181</td>
</tr>
<tr>
<td>16</td>
<td>0.590</td>
<td>41</td>
<td>0.389</td>
<td>82</td>
<td>0.280</td>
<td>210</td>
<td>0.177</td>
</tr>
<tr>
<td>17</td>
<td>0.575</td>
<td>42</td>
<td>0.384</td>
<td>84</td>
<td>0.276</td>
<td>220</td>
<td>0.173</td>
</tr>
<tr>
<td>18</td>
<td>0.561</td>
<td>43</td>
<td>0.380</td>
<td>86</td>
<td>0.273</td>
<td>230</td>
<td>0.169</td>
</tr>
<tr>
<td>19</td>
<td>0.549</td>
<td>44</td>
<td>0.376</td>
<td>88</td>
<td>0.270</td>
<td>240</td>
<td>0.165</td>
</tr>
<tr>
<td>20</td>
<td>0.537</td>
<td>45</td>
<td>0.372</td>
<td>90</td>
<td>0.267</td>
<td>250</td>
<td>0.161</td>
</tr>
<tr>
<td>21</td>
<td>0.526</td>
<td>46</td>
<td>0.368</td>
<td>92</td>
<td>0.264</td>
<td>260</td>
<td>0.157</td>
</tr>
<tr>
<td>22</td>
<td>0.515</td>
<td>47</td>
<td>0.365</td>
<td>94</td>
<td>0.262</td>
<td>270</td>
<td>0.153</td>
</tr>
<tr>
<td>23</td>
<td>0.505</td>
<td>48</td>
<td>0.361</td>
<td>96</td>
<td>0.259</td>
<td>280</td>
<td>0.149</td>
</tr>
<tr>
<td>24</td>
<td>0.496</td>
<td>49</td>
<td>0.358</td>
<td>98</td>
<td>0.256</td>
<td>290</td>
<td>0.145</td>
</tr>
<tr>
<td>25</td>
<td>0.487</td>
<td>50</td>
<td>0.354</td>
<td>100</td>
<td>0.254</td>
<td>300</td>
<td>0.141</td>
</tr>
</tbody>
</table>

"Teaching Manual on "Statistical Methods for Applied Sciences" - Ravi R Saxena, M.L. Lakhera and Roshan Bhardwaj, Department of Agricultural Statistics and SS (L), IGKV, Raipur – 492 012 (Chhattisgarh)"