Also business schools cannot teach people as to how to find a well paid job, even the highest grade candidates will struggle in gaining employment. People skills not just the grades matter.

CHAPTER III

EVALUATION OF TESTING COMPONENTS

3.1 INTRODUCTION

In this chapter an attempt have been made to evaluate the statistical tools applied in the analysis of the present study. The statistical tools used as the testing components of the present study are conventional percentage analysis, arithmetic mean, standard deviation, chi-square test, KMO Measures, factor analysis, rank correlation coefficient and t-test.

3.2 CONVENTIONAL PERCENTAGE ANALYSIS

After collecting information from the respondents, the data were analyzed according to the objectives.
First the data were recorded and data sheets and then fed into the computer. The excel programs were used to compute the percentage of percentages, when and where necessary for the analysis.

3.3 ARITHMETIC MEAN AND STANDARD DEVIATION

Arithmetic Mean

The arithmetic mean is the best known and the most widely used measures of central tendency in statistical work. It is the quotient that results when the sum of all the observation is divided by the number of observations. It is generally denoted by the symbol of $\overline{x}$, read as X bar. We shall abbreviate this term and call it simply mean. Since the geometric mean and the harmonic mean are also measures of central tendency, it must be understood that whenever we speak of mean, we are referring to the arithmetic mean.

Calculation of Mean – Individual Observations

In case of individual observations, mean is obtained by adding the different values of a variable and dividing the sum by the number of observations. Expressed in the formula form:
The sum of all values

\[ \text{Mean} = \frac{\sum_{i=1}^{N} X_i}{N} \]

Symbolically, if \(X_1, X_2, X_3, \ldots, X_N\) are the values of a variable, the mean is computed by the formula:

\[ \overline{X} = \frac{X_1 + X_2 + X_3 + \ldots + X_N}{N} \]

(The notation \(\Sigma\) is read as sigma)

Where,

\(\Sigma = \) the sum of

= The mean of values

\(X_i = \) Values of the variable

\(N = \) Number of values
The standard deviation is the most important, the most reliable and the most widely used measure of dispersion. Like the mean deviation, the standard deviation is also based upon the deviations of the values from mean but it involves a different method of averaging the deviations. Here deviations are squared to make them positive. It is the square-root of the quotient obtained by dividing the sum of the squares of deviations of items from the arithmetic mean by the number of observations.

To quote Yule and Kendell, “The standard deviation is the square-root of the arithmetic mean of the squares of all the deviations, deviations being measured from the arithmetic mean of the observations”. The standard deviation is always measured from the mean the sum of squared-deviations from any other value.

The main drawback of mean deviation is that it ignores algebraic signs but standard deviations have overcome this weakness by squaring the deviations and the process of square root employed in standard deviation neutralizes the squared deviations because the square-root is the anti process of squares. Thus, the standard deviation is simply an average of the deviations.
Why standard deviation is preferred over Mean deviation

The standard deviation is regarded as the best measure of dispersion when compared to mean deviation due to the following reasons.

1. In the calculation of mean deviation, the algebraic signs are ignored which is mathematically illogical. Whereas in the calculation of standard deviation, these signs are considered.

2. The standard deviation is amendable to algebraic treatment. It implies that we can find the combined standard deviation of two or more series. But it is not possible to compute the combined mean deviation.

3. The standard deviation is used in the analysis of statistical series such as in correlation, skewness, sampling or the analysis of variance.

The standard deviation is represented by the Greek letter $\sigma$ (pronounced as sigma) and is computed with the aid of the following formula:

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

The above formula may also be written as

$$\text{Standard deviation} = \sqrt{\frac{\text{Sum of squared deviations from mean}}{\text{Number of Observations}}}$$
\[ \sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}} \]

Where,

\[ x_i \] = \[ x_i \] - \[ \bar{x} \], dev the mean of values from mean

For a frequency distribution:

\[ \sigma = \sqrt{\frac{\sum_{i=1}^{N} f (x_i - \bar{x})^2}{N}} \]

\[ = \sqrt{\frac{\sum_{i=1}^{N} f x_i^2}{N}} \]

The standard deviation is also called the root-mean square because it is the square-root of the mean of the squared deviation.

3.4 CHI-SQUARE TEST

The main use of \( \chi^2 \) is in testing goodness of fit. This is testing of data to see how the theory fits an experiment. Whenever an ideal frequency curve whether normal or some other type is fitted to the data, our main concern is to analyze
whether the difference between observed and expected frequencies is significant or not. This is done by comparing the calculated value of $\chi^2$ with the table value of $\chi^2$ for given degrees of freedom at a certain level of significance.

In case the calculated value of $\chi^2$ is greater than the table value of $\chi^2$, the formulated null hypothesis is rejected. In such a situation, theory does not fit an experiment i.e., the fit is poor. On the contrary, when the calculated value of $\chi^2$ is less than the table value of $\chi^2$, the formulated null hypothesis is accepted. In such a case, the theory fits an experiment and the fit are best.

To test the goodness of fit, the expected frequencies may be obtained from various sources.

**Procedure of Testing a Hypothesis**

The procedure of testing a hypothesis with the aid of $X^2$ – test is illustrated below:

1. Formulate the null hypothesis.

2. Select the level of significance. Generally 5 % and 1 % level of significance are employed.
3. Compute the expected frequencies on the basis of formulated null hypothesis as highlighted in the succeeding pages.

4. If any of the expected frequencies of step (3) are less than 5, combine those cells with each other or with other cell frequencies so that each expected frequency is at least 5.

5. Compute $\chi^2$ by the expression:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

6. Find the degrees of freedom. If the original number of cells is $K$ and no cell combined, the degrees of freedom are computed by the formula:

$$\text{D.F.} = K - 1$$

In case, one cell is combined with others, there is $K'$ cells and the number of degrees of freedom is determined by the expression:

$$\text{D.F.} = K - 1$$

Thus, the number of degrees of freedom is one less than the number of cells after any necessary combinations of cells.
7. Compare the computed value of $\chi^2$ with the table value of $\chi^2$ at a specified level of significance for given degrees of freedom and draw conclusion.

**Conditions in the Application of the $X^2$ Test**

For testing the significance with the aid of $\chi^2$ – test, the following precautions should be observed:

1. The total number of observations (N) must be reasonably large. Otherwise the individual values (X’s) are not normally distributed. It is very difficult to draw a line that constitutes largeness. But as an arbitrary figure, we may see that ‘N’ should be at least 50, however few the cells.

2. No theoretical frequency should be small. Here, again, it is difficult to say what constraints smallness. 5 should be regarded as the very minimum but 10 are much better. Whenever, theoretical frequencies are less than 5, they should be amalgamated.

3. The constraints must be linear.

4. For applying $\chi^2$ test, absolute frequencies rather than relative frequencies should be used.
To these four conditions Yule and Kendall in their book “An Introduction to the Theory of Statistics: added the following remarks, which must be considered while applying $\chi^2$-test.

(a) The $\chi^2$ – test enables us to compute the probability of getting, on a random sample, a value of $\chi^2$ equal to or higher than the actual value. If the probability is small we are justified in suspecting a significant divergence between theory and experiment.

(b) Not only small values of probability lead to suspect the hypothesis but a value of probability very near to unity may also do so.

3.5 KMO – MEASURES

Before extracting the factors to test the appropriateness of the factor model, Bartlett’s test of sphericity was used to test the null hypothesis that the variables are intercorrelated in population. The test statistics of sphericity is
based on a chi-square transformation of the determinant of the correlation matrix.

Another useful statistic is the Kaiser – Meyer Oklin (KMO) test of sampling adequacy\textsuperscript{100}. The small value of the KMO statistic indicates that the correlation between parts of the variable cannot be explained by other variables and that factor analysis may not be appropriate. Generally, a value greater than 0.5 is desirable.

The correlation matrix was examined carefully and the two tests namely Bartlett’s test of Sarasota and Kaiser-Meyer Oklin test was undertaken to test if it was judicious to proceed with factor analysis in the present study. The computed results are given in Table 3.1.

\begin{table}[h]
\centering
\begin{tabular}{|l|c|}
\hline
\textbf{MEASURES OF SAMPLING INADEQUACIES} \\
\hline
\end{tabular}
\end{table}

\textsuperscript{100} Marjorie A.Pett, Nancy R.Lackey and John J.Sullivan, (2003), Marketing Sense of Factor Analysis, Sage Publications, New Delhi, pp. 73-78.
<table>
<thead>
<tr>
<th>Measures</th>
<th>Estimated Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kaiser – Meyer Oklin Measures of Sampling Adequacy</td>
<td>0.7862</td>
</tr>
<tr>
<td>Bartlett’s Test of Sphercity</td>
<td></td>
</tr>
<tr>
<td>Appropriate chi-square</td>
<td>2617.42</td>
</tr>
<tr>
<td>Degrees of freedom</td>
<td>861</td>
</tr>
<tr>
<td>Significance</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

From Table 3.1 it has been observed that the Bartlett’s test was significant with \( P = 0.000 \), being less than 0.05. Sampling adequacy measured using the Kaiser – Mayer Oklin (KMO) of 0.7862 was taken as acceptable. Thus the factor analysis may be considered an appropriate technique for analyzing the data.

### 3.6 FACTOR ANALYSIS

The technique adopted for extracting the factors from the complex mass of variables is “Factor Analysis”. The principal factor analysis method is mathematically satisfying because it yields a mathematically unique solution to a factor problem. Its major solution feature is the extraction of the maximum amount of variance as each factor is calculated. In other words, the first factor
extracts the most variance, the second the next most variance, the third the next most variance and so on.\(^{101}\)

The principal factor method actually involves the solution of simultaneous linear equations. The roots obtained from the solution are called Eigen values. Eigen factors are also obtained after suitable transformation and they become the factor loadings. The fictitious ‘R’ matrix was solved in this manner, yielding the factor matrix to be given.

Most of the analytic methods produce results in a form that is difficult or impossible to interpret. Thurston argued that it was necessary to rotate factor matrices if one wanted to interpret them adequately.\(^{102}\)

He pointed out that original factor matrices are arbitrary in the sense that an infinite number of reference frames (axes) can be found to reproduce any given ‘R’ matrix.\(^{103}\)

The extracting factors of the observed correlation matrix arbitrarily located the reference axes in a different position. These reference axes were rotated by


\(^{103}\) Ibid., P.93.
employing Varimax method. From the arbitrary location to some position which facilitated the interpretation of the factors in conformity with the purpose of the study. A major goal of the rotation was to obtain meaningful factors that were as much consistent as possible from analysis to analysis. Each variable finds itself positioned with that factor, with which it is highly correlated which means that factor in which there is the highest loading.

The criteria used for the building up of factors were such that:

1. The ratios (variables) with the highest rotated factor loading under particular factor.
2. The ratios (variables) allocated to one particular factor are independent of those allocated to other factors; and
3. The factors are formed in such a way that the cumulative percentage of the total variance attributable to each successive factor is maximum.

The common factors extracted by the above criteria were independent of each other. They were identified by providing a reasonable interpretation of their

---


underlying significance. Thus the factors (ratios) that tend to influence the growth of the pharmaceutical industry are identified.

**Assumption – 1**

Certain basic assumptions are taken for granted while factor analyzing the inter-correlated variables. If the two ratios (variables) are correlated with each other, it may be incidental to the fact that both are correlated with a third variable or a set of variables.

**Assumption – 2**

The second assumption of factor analysis is that the total variance of a ratio has three components; common, specific and error variables. This is expressed by the equation.

\[ V_t = V_{co} + V_{sp} + V_e \]

Where,

- \( V_t \) - Total variance of a ratio (variable)
- \( V_{co} \) - Common factor variance, or the variance that two or more Ratios (variables) share in common
\( V_{sp} \) - Specific variance, or the variance of the variable that is not
with any other ratio (variable), that is the variance of that
ratio (variable) and no other;

\( V_e \) - Error variance

\( V_{cp} \) can be written as \( a_{i1}^2 + a_{i2}^2 + a_{i3}^2 + ...... + a_{ik}^2 \)

The value of \( a_{i1}^2 \), \( a_{i2}^2 \), ...... \( a_{ik}^2 \) are called factor loading which represents the
degree of correlation of the it ratio (variable) with each other. The sum of squares
of the factor loading of its ratio (variable) is community and common factor
variance of that ratio (variable). That is,

\[
    h_i^2 = a_{i1}^2 + a_{i2}^2 + a_{i3}^2 + ...... + a_{ik}^2;
    \text{But } V_{co} = h_i^2
\]

To find a proportional representation;

\[
    \frac{V_i}{V_s} = \frac{V_{co}}{V_s} + \frac{V_{sp}}{V_s} + \frac{V_c}{V_s}
\]

The first term on the right hand side of the above equation is associated
with the rapid or common factor variance of the ratio (variable), and the first two
terms on the right are associated with the reliability of the ratio (variable), labeled as ours.107

Assumption – 3

The third assumption is that the product of the unrotated factor matrix and its transpose is equal to the observed correlation matrix. In notation it can be written as \( R = FF' \), where \( F \) is the unrotated factor matrix and \( R \) is the observed correlation matrix. This is called the basic equation of the factor analysis.

Methods of Factor Analysis

There are a number of methods available for factor analyzing108. But the principal factor method with orthogonal varimax rotation is most used and widely available in factor analytic computer programming. Further orthogonal rotations maintain the independence of factors; that is, the angles between the axes are kept at 90 degrees. One of the final outcomes of a factor analysis is called the Rotated Factor Matrix, a table of co-efficient that expresses the ratios between the variables in the table is called factor loadings. The factor loadings range from -1 through 0 to +1. They


express the correlations between the variables (ratios) and the factors. The sums of squares of the factor loadings of a variable are called communities or $h^2$. The communality of a variable is its common factor variance. It would give us clues to its common factor variance. It would give us clues to its nature by telling us which other tests share the same common factor variance and which do not.\textsuperscript{109} The variables (ratios) with factor loadings of 0.70 or greater have been considered as significant variables (ratios). This limit was chosen because it was judged that variables with less than 50 per cent common variation with the rotated factor pattern (which equals the square of factor loading times 100) were too weak to report.

3.7 COEFFICIENT OF RANK CORRELATION

The Karl Pearson’s coefficient of correlation is based upon the assumption that the population from which variables are taken is normal. When the population departs from normality, there arises a need for a measure of correlation that involves no assumption about the normality of the population.

A statistic to measure the degree of correlation between two such variables was developed by Charles Edward Spearman in 1904 and is called the coefficient of rank correlation. It is denoted by Greek letter ‘ρ’ (Pronounced as Rho).

In many situations, we cannot measure the values but we can arrange the individuals in order of merit. For example a teacher may rank his students in order of merit without assigning marks to each student. Similarly a judge in a beauty contest ranks the contestants in order of merit by giving any indication how much better the winner is than the runners up. Ranking is mostly used in those fields where the measurement of characteristic being judged is not possible due to shortage of time, lack of money or unreasonably defined units. Sometimes ranking method is also employed in those cases where measurements can be made, to reduce the labor of computation or to get the results quickly.
Spearman’s rank correlation coefficient is computed by the following formula:

\[ \rho = 1 - \frac{6 \sum d^2}{N (N^2 - 1)} \]

Where,

‘d’ stands for difference in ranks assigned to the same individual.

N stands for the number of paired individuals.

3.8 T – TEST

Test of Significance of Difference between Two Means

A problem that arises in statistical investigations is that of determining whether two samples have been drawn from the same parent population or from populations which are alike in respect of some stated parameter. In this section, we discuss the procedure employed in testing the significance of difference between two means.

Suppose, two samples consisting of \( N_1 \) and \( N_2 \) items are drawn from the two populations; \( \bar{x}_1 \) and \( \bar{x}_2 \) are the means and \( S_1 \) and \( S_2 \) are the standard
deviations respectively. Now the problem crops up: whether the difference between two means is significant or not significant i.e. Solely as a result of sampling fluctuations. In such cases, we start with the null hypothesis that the true difference between two means is zero \((D = 0)\). If we draw successive pairs of samples from the same population, we get a series of values of ‘D’: some plus and some minus. The sampling distribution of \(D\) (differences) would be distributed in accordance with normal law about mean zero. Therefore, we have to find the standard deviation of ‘D’ in order to know variations in the \(D\)’s (differences) from sample to sample. It is often called the standard error of difference between two means and is computed by the formula of type.

\[
S.E \ \bar{x}_1 - \bar{x}_2 = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}
\]

In the above formula, we have supposed that samples come from two populations with known standard deviations \(\sigma_1\) and \(\sigma_2\) respectively. In case the two samples come from the same population with known standard deviation \((\sigma_0)\), the above formula may be written as:

\[
S.E \ \bar{x}_1 - \bar{x}_2 = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_2}}
\]
In many situations, we do not know the standard deviations of the populations or population from which samples have been drawn. In such cases, we can substitute the standard deviations of samples in place of population standard deviations. The above formula may be stated as:

\[
S.E\ \overline{X_1} - \overline{X_2} = \sqrt{\frac{S^2_1}{N_1} + \frac{S^2_2}{N_2}}
\]

Lastly we have to determine the critical ratio ‘T’ for analyzing whether the difference between two means is significant or insignificant. The critical ratio ‘T’ is defined by the relation:

\[
T = \frac{\overline{X_1} - \overline{X_2}}{S.E\ \overline{X_1} - \overline{X_2}}
\]

If the value of ‘T’ falls in the acceptance region, the null hypothesis is accepted. On the contrary, if the value of ‘T’ falls in rejection region, the null hypothesis is rejected and some alternative hypothesis is accepted.