CHAPTER 2

STATISTICAL TOOLS USED FOR THE STUDY

To understand the behaviour of the investors with reference to the various variables at work, a number of statistical tools are available. Of these, the most prominent tools are:

i) Factor Analysis
ii) Analysis of Variance
iii) Discriminant Analysis
iv) Cluster Analysis

A brief description of the statistical tools is given in this chapter.

2.1 FACTOR ANALYSIS – AN OVERVIEW

A number of methods exists for solving this problem as given by Rummel\(^1\), but one of the most popular is the method of Principal Components. Principal components analysis is performed in order to simplify description of a set of inter-related variables. In principal components analysis, the variables are treated equally. That is, they are not divided into dependent and independent variables, as in regression analysis. The technique can be summarised as a method of transforming the original variables into new, uncorrelated variables. The new variables

are called the principal components. Each principal component is a linear combination of the original variables. One measure of the amount of information conveyed by each principal component is its variance. For this reason, the principal components are arranged in order of decreasing variance. Thus, the most informative principal component is the first, and the least informative is the last. (A variable with zero variance does not distinguish between the members of the population).

The researcher may wish to reduce the dimensionality of the problem. That is, reduce the number of variables without losing much of the information. This objective can be achieved by choosing to analyse only the first few principal components. The principal components not analysed convey only a small amount of information since their variances are small. This technique is attractive for another reason, namely, that the principal components are not intercorrelated. Thus instead of analysing a large number of original variables with complex inter-relationships, the researcher can analyse a small number of uncorrelated principal components. The selected principal components may also be used to test for their normality. If the principal components are not normally distributed, then neither are the original variables. Another use of the principal components is to search for outliers. A histogram of each of the
principal components can identify those individuals with very large or very small values; these values are candidates for outliers or blunders.

In regression analysis it is sometimes useful to obtain the first few principal components corresponding to the X variables and then perform the regression on the selected components. This tactic is useful for overcoming the problem of multicollinearity since the principal components are uncorrelated. Principal components analysis can also be viewed as a step toward factor analysis.

Principal components analysis is considered to be an exploratory technique that may be useful in gaining a better understanding of the inter-relationships among the variables. 

Factor analysis is similar to Principal components analysis in that it is a technique for examining the inter-relationships among a set of variables. Both these techniques differ from regression analysis in that we do not have a dependent variable to be explained by a set of independent variables. However, principal components analysis and factor analysis also differ from each other. In principal components analysis, the major objective is to select a number of components that explain as much of the total variance as possible. The values of the principal components for a given individual are relatively simple to compute and interpret. On the other hand, the factors obtained in factor analysis are selected mainly to
explain the inter-relationships among the original variables. Ideally, the number of factors expected is known in advance. The major emphasis is placed on obtaining easily understandable factors that convey the essential information contained in the original set of variables.

In factor analysis, we begin with a set of variables $X_1$, $X_2$, ..., $X_n$. These variables are usually standardized by the computer program so that their variances are each equal to one and their covariances are the correlation coefficients.

The object of the Factor analysis is to represent each of these variables as a linear combination of a smaller set of common factors plus a factor unique to each of the response variables.

We express this representation as

$$x_i = l_{i1}F_1 + l_{i2}F_2 + ... + l_{im}F_m + e_i$$

$$x_p = l_{p1}F_1 + l_{p2}F_2 + ... + l_{pm}F_m + e_p$$

$$x_p = l_{p1}F_1 + l_{p2}F_2 + ... + l_{pm}F_m + e_p$$

The following are the assumptions under Factor analysis.

1. $m$ is the number of common factors (this number is much smaller than $P$).
2. \( F_1, F_2, \ldots, F_m \) are the common factors. These factors are assumed to have zero means and unit variances.

3. \( l_{ij} \) is the coefficient of \( F_i \) in the linear combination describing \( x_j \). This term is called the loading of the \( j \)th variable on the \( i \)th common factor.

4. \( e_1, e_2, \ldots, e_p \) are unique factors, each relating to one of the original variables.

The above equations and assumptions constitute the so-called factor model. Thus each of the response variables is composed of a part due to the common factors and a part due to its own unique factor. The part due to the common factors is assumed to be a linear combination of these factors.

Suppose that \( x_1, x_2, x_3, x_4, x_5 \) are the standardized scores of an individual on five tests. If \( m = 2 \) the following model is assumed:

\[
x_1 = l_{11}F_1 + l_{12}F_2 + e_1
\]

\[
x_2 = l_{21}F_1 + l_{22}F_2 + e_2
\]

\[
x_5 = l_{51}F_1 + l_{52}F_2 + e_5
\]
Each of the five scores consists of two parts: a part due to the common factors $F_1$ and $F_2$ and a part due to the unique factor for that test. The common factors $F_1$ and $F_2$ might be considered the individual's verbal and quantitative abilities, respectively. The unique factor includes all other effects that keep the common factors from completely defining a particular $x_i$.

The Factor model is the mirror image of the principal components model, where each principal component is expressed as a linear combination of the variables. Also, the number of principal components is equal to the number of original variables. On the other hand, in Factor analysis, the number of factors chosen are smaller than the number of response variables. Ideally, the number of factors should be known in advance, although this is often not the case. However, it is possible to allow the data to determine this number.

The factor model, by breaking each response variable $x_i$ into two parts, also breaks the variance of $x_i$ into two parts. Since $x_i$ is standardized, its variance is one. This variance is one composed of the following two parts.

1. The communality, i.e., the part of the variance that is due to the common factors.
2. The *specificity*, i.e., the part of the variance that is due to the unique factor $e_i$.

Denoting the *communality* of $x_i$ by $h_i^2$ and the *specificity* by $u_i^2$, we can write the variance of $x_i$ as $\text{var } x_i = 1 = h_i^2 + u_i^2$. In other words, the variance of $x_i$ equals the communality plus the specificity.

The numerical aspects of factor analysis are concerned with finding estimates of the *factor loadings* ($l_{ih}$) and the *communalities* ($h_i^2$). There are many ways available to numerically solve for these quantities. The solution process is called *initial factor extraction*. Once a set of initial factors is obtained, the next major step in the analysis is to obtain new factors, called the rotated factors, in order to improve the interpretation.

2.2 ANALYSIS OF VARIANCE – A DESCRIPTION

The *analysis of variance* frequently referred to by the contraction ANOVA is a statistical technique specially designed to test whether the means of more than two quantitative populations are equal. Basically, it consists of classifying and cross-classifying statistical results and testing whether the means of a specified classification differ significantly. In this way, it is determined whether the given classification is important in affecting the results. The analysis of variance test is not intended to serve the ultimate purpose of testing for the significance of the difference
between the sample variances; rather its purpose is to test for the significance of the differences among sample means. They do this via the mechanism of the F-test for testing for the significance of the difference between two variances, but the test is so designed that the variances being compared are different only if the means under consideration are not homogeneous. In this way, significant values of \( F \) indicate that the means are significantly different from one another.

Following are the assumptions in Analysis of Variance:

(i) Normality,

(ii) Homogeneity, and

(iii) Independence of error

One-way Classification:

In one-way classification, the data are classified according to only one criterion. The null hypothesis is:

\[
H_0: \mu_1 = \mu_2 = \mu_3 \ldots = \mu_k
\]

That is, the arithmetic mean of populations from which the \( k \) samples were randomly drawn are equal to one another.
The steps in carrying out the analysis are:

(i) The variance between the samples (groups) measures the difference between the sample mean of each group and the overall mean weighted by the number of observations in each group. The sum of squares between samples is denoted by SSC. For calculating variance between the samples, the total of the square of the deviations of the means of various samples from the grand average should be divided by the degrees of freedom. The Degrees of freedom will be one less than the number of samples.

(ii) The variance (sum of squares) within samples measures those inter-sample differences due to chance only. It is denoted by SSE. The variance within samples, measures variability around the mean of each group. Since the variability is not affected by group differences, it can be considered a measure of the random variation of values within a group. For calculating the variance within the samples, the total sum of squares of the deviation of various items from the mean values of the respective samples should be divided by the degrees of freedom.

(iii) The ratio of $F$ has to be calculated:

$$ F = \frac{\text{Between-column variance}}{\text{Within-column variance}} $$
\[ F = \frac{S_1^2}{S_2^2} \]

The F-distribution (named after the famous Statistician R.A. Fisher) measures the ratio of the variance between groups to the variance within groups. The variance between the sample means is the numerator and the variance within the sample means is the denominator. If there is no real difference from group to group, any sample difference will be explainable by random variation and the variance between groups should be close to the variance within groups. However, if there is a real difference between the groups, the variance between groups will be significantly larger than the variance within groups.

(iv) The calculated value of F should be compared with the Table value of F for the degrees of freedom at certain level (generally at 5% level of significance). If the calculated value of F is greater than the table value, it is concluded that the difference in sample means is significant, i.e., it could not have arisen due to fluctuations of simple sampling or, in other words, the samples do not come from the sample population. On the other hand, if the calculated value of F is less than the table value, the difference is not significant and has arisen due to fluctuation of simple sampling.
### ANOVA Table

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sum of Squares</th>
<th>Degrees of Freedom</th>
<th>Mean Square</th>
<th>Variance Ratio of F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Samples</td>
<td>SSC</td>
<td>( V_1 = c-1 )</td>
<td>MSC = SSC/(c-1)</td>
<td></td>
</tr>
<tr>
<td>Within Samples</td>
<td>SSE</td>
<td>( V_2 = n-c )</td>
<td>MSE = SSE/(n-c)</td>
<td>MSC/MSE</td>
</tr>
<tr>
<td>Total</td>
<td>SST</td>
<td>( N - 1 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **SST** = Total sum of squares of variations
- **SSC** = Sum of squares between the samples (Columns)
- **SSE** = Sum of squares within samples (rows)
- **MSC** = Mean sum of squares between samples
- **MSE** = Mean sum of squares within samples.

**Rationale of the Test:**

The variation within the samples, i.e., the variation of the individual observations within the samples from their own individual sample means, measures the influence of the chance forces which cause the individual observations to vary from one another. However, the variation of the \( K \) sample means from the grand sample means of all the samples taken together, i.e., the variation between the samples means reflects not only the effect of these same chance forces, but also the effect of the forces, if
any, which cause the various sample means to differ from one another. Thus, if there is any such force, i.e., if the hypothesis is not true, the variation between the sample means will tend to be larger than the variation within the samples.

2.3 DISCRIMINANT ANALYSIS – THEORETICAL ASPECTS

Discriminant analysis deals with the problem of distinguishing (discriminating) between different populations on the basis of observation on several variables. In other words, Discriminant analysis techniques are used to classify individuals into one of two or more alternative groups (or populations) on the basis of a set of measurements.

The Statistical problem is to develop a rule, or discriminant functions, based on the measurements obtained on each of these individuals, that will help us to assign some new individual to the correct population when it is not known from which of the two populations the individual comes.

The data obtained on each individual will invariably consist of observed values of a set of mutually correlated random variables, and the presence of these intercorrelations will necessitate consideration of the variables together rather than one at a time. The general approach, as with regression analysis, is to construct in some optimal way a linear combination of these variables that is then used for classification. This transforms the basic problem from a complex multivariable one to an easier-to-handle univariable one, and the assignment of an individual to one of the two populations is then based simply on the value of the linear combination for that particular individual. The statistical manipulations associated with constructing a good (i.e. optimal) linear combination and with developing related methodology differ from those of regression analysis and are classed under the general heading discriminant analysis. Discriminant analysis requires multivariate normality of the independent variables. The basic strategy in discriminant analysis is to form a linear combination of the variables.

\[ Y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \]

And then to assign a new individual to either group 1 or group 2 on the basis of the value of \( Y \) obtained for that new individual.
More specifically, for any individual in the group of \( n1 + n2 \) individuals available, one can calculate (if \( \beta \)'s is known) the value of \( Y \) for that individual. In particular, for the \( k \)th individual selected from population \( i \), the associated value of \( Y \) is

\[
Y_{ik} = \beta_1 x_{1k} + \beta_2 x_{2k} + \ldots + \beta_p x_{pk}
\]

For each individual, one can convert or transform from a set of \( P \) variable values to a single univariate score. Once this transformation is made, the problem is to distinguish between the two populations on the basis of values \( \langle L_{11}, L_{12}, \ldots, L_{1n1} \rangle \) from population 1 and values \( \langle L_{12}, L_{22}, \ldots, L_{2n2} \rangle \) from population 2.

Statisticians have formulated different ways of performing and evaluating discriminant functions analysis. One method of evaluating the results is Classification functions.

Based on this classification functions, an individual is classified as coming from one of the \( k \) population. If the number of populations \( k=2 \), then a single discriminant function is constructed and depending on the value of the function an individual is classified as either belonging to population 1 or 2. If \( k=3 \), two populations are needed to classify an observation. In general, if there are \( k \) population, then \( k-1 \) functions are required to classify an observation into either one of these \( k \)
populations. These \( k-1 \) discriminant functions are independent among themselves.

The power of the discriminant function constructed based on the past data can be checked based on various criteria. One such criteria is to check whether the discriminant functions correctly classify the existing past observation. That is, the populations from which the past observations are obtained is known. After constructing the discriminant functions based on these observations, these observations are classified based on the discriminant functions. This can be called as predicted population. How can an individual be classified into one of these populations (Predicted Population).

It is helpful to think of associating with each individual a score that is the value of \( Y \) based on that particular individual's set of observed variable values. Then, in order to use \( Y \) for allocating individuals to one of the two groups, a cut off point is required, such that an individual is assigned to one group if his score exceeds the cut off point and to the other if it does not. The cut off point is just nothing but the grand mean of scores for the samples of different groups.

\[
C = \frac{\bar{x}_1 + \bar{x}_2 \ldots \ldots \bar{x}_n}{n}
\]

If the observation belongs to population 1 and its predicted population is also 1, then it is correctly classified, otherwise there is a
misclassification. If the percentage of correct classification is larger and misclassification is small, it is said that the discriminant function really discriminates the populations. If the discriminant function is good in discriminating between populations then any future observations from an unknown population can be classified as coming from any one of the k populations.

The power of the discriminant function also depends on the set of measurements or variables based on which it is constructed. If there is no variation between k population for these variables then the discriminant function constructed using these variables may not be useful. Hence there is a need for selecting the variables for discriminant analysis. This can be done using stepwise discriminant analysis, which selects the variables, which are really helpful in discriminating these k populations.

2.4 CLUSTER ANALYSIS - CONCEPTUAL FRAME WORK

Cluster analysis is a technique for grouping individuals or objects into unknown groups. It differs from other methods of classification, such
as discriminant analysis, in that, in cluster analysis, the number and characteristics of the groups are to be derived from the data and are not usually known prior to the analysis.

A helpful technique for a moderate number of variables is a Profile diagram. To plot a profile of an individual case in the sample, the researcher customarily first standardizes the data by subtracting the mean and dividing the standard deviation for each variable. A profile diagram lists the variables along the horizontal axis and the standardized value scale along the vertical axis. Each point on the graph indicates the value of the corresponding variable. The points are connected in order to facilitate the visual interpretation. A preliminary clustering procedure is to graph the profiles of all cases on the same diagram.

Distance measures

The analytical methods require defining some measures of closeness of similarity of two observations. The converse of similarity is Distance. Many analytical techniques are sensitive to outliers. Some preliminary checking for outliers and blunders is advisable.
The most commonly used distance is the Euclidian Distance. In two dimensions, suppose that two points have co-ordinates \((X_{11}, X_{21})\) and \((X_{12}, X_{22})\) respectively, then the Euclidian distance between the two points is defined as

\[
\text{Distance} = \sqrt{(X_{11} - X_{12})^2 + (X_{21} - X_{22})^2}
\]

For \(P\) variables, the Euclidian distance is the square root of the sum of the squared differences between the co-ordinates of each variable for the two observations. The Euclidian distance is not invariant to changes in scale and the results can change appreciably by simply changing the units of measurement. In computer program output, the distances between all possible pairs of points are usually summarized in the form of matrix. Since the square root operation does not change the order of how close the points are to each other, some programs use the sum of the squared differences instead of the Euclidian distance (i.e. without square root).

Several other definitions of distance exist. Here one commonly used definition, the Mahalanobis distance. The Mahalanobis distance is a generalization of the idea of standardization. The squared Euclidian distance based on standardized variables is the sum of the squared
differences, each divided by the appropriate variance. When the variables are correlated, a distance can be defined to take this correlation into account. For two variables \(X_1\) and \(X_2\), suppose the sample variances are \(S_1^2\) and \(S_2^2\), respectively, and that the correlation is \(r\). The squared Euclidean distance based on the original value is

\[
(\text{Euclidean distance})^2 = (X_{11} - X_{22})^2 + (X_{21} - X_{22})^2
\]

The same quantity based on standardized variables is

\[
(\text{Standardized Euclidean distance})^2 = \frac{(X_{11} - X_{12})^2}{S_1^2} + \frac{(X_{21} - X_{22})^2}{S_2^2}
\]

For more than two variables, the Mahalanobis distance is easily defined in terms of vectors and matrices.

**Hierarchical clustering**

Hierarchical methods can be either Agglomerative or Divisive. In the agglomerative methods, one begins with \(N\) clusters; each observation constitutes its own cluster. In successive steps, the two closest clusters are combined thus reducing the number of clusters by one in each step. In the final step all observations are grouped into one cluster. **Divisive** methods begins with one cluster containing all of the observations. In successive steps, the cases that are most dissimilar to
the remaining ones are split off. Most of the commonly used programs are of the agglomerative type.

The centroid procedure is a widely-used example of agglomerative methods. In the centroid method, the distance between two clusters is defined as the distance between the group centroids (the centroid is the point whose co-ordinates are the means of all the observations in the cluster). If a cluster has one observation, then the centroid is the observation itself. The process proceeds by combining groups according to the distance between their centroids, the groups with the shortest distance being combined first.

When the number of variables is more than two, a very good device called the Dendrogram or Tree Graph has been incorporated into packaged computer program to summarize the clustering at successive steps.

In Tree graph, the number of clusters is printed on the vertical axis instead of the step number. The order of observations on the horizontal axis is helpful in indicating which observations are sufficiently similar to be combined in the early steps.
An important problem is how to select the number of clusters. No standard objective procedure exists for making the selection. The distances between clusters at successive steps may serve as a guide. The researcher can stop when this distance exceeds a specified value or when the successive differences in distances between steps make a sudden jump.