1. HYPOTHESIS

Hypothesis is a pivot of all kinds of researches. A research without hypothesis is just like a navigator in the sea without compass. After defining the research problem formulation of hypothesis is a fundamental step from where the researcher moves into successive stages of research process to arrive at the findings. Hypothesis is to be formulated in the light of aims, objectives, scope and limitations of the research study. It will be empirically tested and its findings may either reject or accept the hypothesis. Keen observation, creative thinking, hunch, wisdom, imagination, intuition, vision, insight, sound judgment and through knowledge about the phenomenon, situation and related facts and variables are of great significance in the formulation of reasonable hypotheses. The formulation of hypothesis play important role in the growth of knowledge, concept models and theories, which will be useful for the policy makers of all fields to take a correct decision in the backdrop of research findings.

The term 'hypothesis' is derived from the ancient Greek word, 'hypotithemi' that means 'to put under' or 'to suppose'. Hypothesis is also a combination of two words 'Hypo + Thesis, where 'Hypo' means tentative or subject to verification and 'Thesis' a statement based on concepts, theories and past experiences about the solution of the problem. The term hypothesis literally means an assumption or a supposition about the state of affairs of a certain thing or phenomena or facts or variable or situation. Thus, "hypothesis is perceived as a proposition or set of propositions set forth as an explanation for occurrence of some specified group of phenomenon either asserted merely as a provisional conjecture to guide some investigation or accepted as highly probable in the light of established facts. Quite often a research hypothesis is a predictive statement, capable of being tested by scientific methods, that relates an independent variable to some dependent variables" (Kothari, 2004).

The view point of various thinkers has been presented as under:-

In the words of Good, Barr and Scates (1936): "Hypothesis is a statement temporarily accepted as true in the light of what is , at the time, known about a phenomenon, and it is employed as a basis for action in the search for new truth. When hypothesis is fully established, it may take the form of facts, principles and theories".

In the words of Lundberg (1951): "Hypotheses is a tentative generalisation, the validity of which is remains to be tested. In the most elementary stage the hypothesis may be any hunch, guess, imaginative idea which become base for further investigation".

In the words of Best (1963): "Hypothesis is a shrewd guess or inference that is formulated and provisionally adopted to explain observed facts or conditions and to guide in further investigation".

In the words of Mouly (1963): "Hypothesis is an assumption whose testability is to be tested on the basis of the compatibility of its implications with empirical evidences and previous knowledge".

In the words of Gopal (1965): "Hypothesis is a tentative solution posed on cursory observation of known and available data and adopted provisionally to explain certain events and to guide in the investigation of others. It is in fact, a possible solution to the problem".

In the words of Theodorson and Theodorson: "A hypotheses is a tentative statement asserting a relationship between certain facts".

In the words of Goode and Hutt: "A proposition which can be put to a test to determine its validity".

In the words of Kerlinger (1973): "A hypothesis is a conjectural statement of the relationship between two or more variables".

In the words of Black and Champion (1976): "A hypotheses is a tentative statement about something the validity of which is usually unknown".

In the words of While Bailey (1978): "Hypothesis is a proposition that is stated in a testable form and that predicts a particular relationship between two or more variables".

In the words of Grinnell (1988): "Hypothesis is written in such a way that it can be proven or disproven by valid or reliable data—it is in order to obtain these data that we perform our study".

In the words of Palmar O Johnson: "A hypothesis in statistics is simply a quantitative statement about a population".

In the words of Webster: "Hypothesis is a tentative assumption made in order to draw out and test its logical or empirical consequences".

On the bases of these definition it can be concluded that hypotheses is a hunch, assumption, assertion, proposition, predictive or tentative statement, an idea about a phenomena or a situation the truth or reality of which is yet to be known through the process of rigorous empirical testing. It becomes the basis of systematic and scientific enquiry which guides the research process in the proper direction. Hence, "hypothesis is a tentative answer to the defined problem, that has to pass through the process of active testing ; which may or may not be correct."

2. CHARACTERISTICS OF HYPOTHESIS

The following the basic characteristics of a hypothesis:-

1. **Valid:** Hypothesis must be valid and related to the phenomena or situation which it is trying to explain.
2. **Pivot of Research:** Hypothesis is backbone of all kinds of researches because the all research activities are designed to verify the hypothesis from 360 degree angle.

3. **Conceptual Clarity:** Hypothesis must be clearly and precisely stated. There should be no ambiguity in the formulation of hypothesis. It means hypothesis should be defined lucidly, should be operationalsed, should be commonly accepted and should be communicable.
4. **Testability:** A hypothesis should be testable and not moral judgement. It should be possible to collect empirical evidences to test the hypothesis. In the words of C. William Emory, "A hypothesis is testable if other deductions can be confirmed or disapproved by observation".
5. **Specificity:** A hypothesis should be specific and explain the expected the relations between variables and the situations under which these relation will hold.
6. **Consistency:** A hypothesis should be logically consistency. Two or more hypothesis logically derived from the same population must not be mutually contradictory.
7. **Objectivity:** Hypothesis should be free from value judgement. In the scientific research the researcher's value system has no place in scientific enquiry.
8. **Simplicity:** Hypothesis should be a simple one requiring fewer assumptions. But the simplicity does not mean vague idea.
9. **Theoretical Relevance:** Hypothesis should be based upon some theoretical foundations. When a research is systematically based upon a body of existence knowledge, only then a genuine contribution is more likely to result in.
10. **Availability of Technique:** Hypothesis should be related to available techniques, otherwise it will not be researchable. Hence, the researcher must ensure that statistical or mathematical techniques are available for testing the proposed hypothesis.
11. **Future Oriented:** Hypothesis is forward looking concept as it is related to future verification not the past facts, information or situations.

3. FUNCTIONS OF THE HYPOTHESIS

Hypothesis serves the following important functions:-

1. **Guide and Direction:** Hypothesis provides the definite base to the scientific investigation. It guides the direction of the research, without which research becomes unfocused.
2. **Source of Data:** It specifies the source of data which is required to test the hypothesis.
3. **Determine Data:** It determine the data required in the investigation. It defines which facts are relevant and which are not. It prevent the blind research and indiscriminate collection of data.
4. **Type of Research:** It suggests which type of research is likely to be most appropriate.
5. **Appropriate Technique:** It determine the most appropriate technique of the research.
6. **Contribute to the Development of Theory:** Hypothesis contributes to the development of body of knowledge and theory. It links theory with present research. When it is proved true then it forms a part of theory.

4. SOURCES OF THE HYPOTHESIS

Hypothesis can be derived from the various sources, which are presented as under:-

1. **Theory:** Theory is one the main sources of the hypothesis. It gives direction to research

what is known and what is to be known. Logical deductions from theory lead to the formulation of new hypothesis.

2. **Observations:** Observation can be one of the important sources of the hypothesis. Random observations during discussions, conversations and reflections on the life of persons also throw light on events and issues.
3. **Analogies:** It is another useful source of the hypothesis. Here, the defined problem is compared with the previous studies or researches and then the hypothesis is formulated in accordance with past experience and present circumstances.
4. **Intuition and Personal Experience:** Personal life experiences also determine researcher's perception and conception. These may in turn guide a person for formulation of hypothesis more quickly.
5. **Findings of the Existing Studies:** Hypothesis may also be formulated from the findings of the existing studies in order to replicate and test them.
6. **Body of Knowledge:** An important source of hypothesis is existing body of knowledge in any particular subject. Where formal theories exist, hypothesis can be deduced. In case hypotheses are rejected, theories would be modified.
7. **Culture:** A particular culture in which researcher is nurtured is also an important source of hypothesis. Western culture emphasises on individualism, democracy, competition and equality so research will revolve around these subjects. Whereas Indian culture emphasises on traditions, collectivism, values, karma, puruswartha and spirituality so research will revolve around these areas.
8. **Continuity of Research:** The continuity of research in a particular itself constitutes an important source of hypothesis. For example research in the area of corporate governance can also provide fertile ground for research in the area of corporate social responsibility and business ethics. The rejection of hypothesis leads to the formulation of new insight in that area.
9. **Folk Wisdom:** Sometimes researcher can get idea of a hypothesis commonly held lay belief like role of caste, religion and region in Indian polity.

■ TYPES OF HYPOTHESIS

reject null hypothesis).

8. TEST OF SIGNIFICANCE

Various tests of significance have been developed to meet various types of requirements in business, economic, social, financial and operational environments etc. These are both parametric and non-parametric test.

1. **Parametric tests** are statistical techniques to test a hypothesis based on certain assumptions about the population. The most important parametric tests are: the z-test, the t-test and the F-test.

They are based upon the following assumptions:-

- The observations or value must be independent, means the selection of one item must not depend upon the selection of any other item.
- The population from which the sample is drawn at a random basis should be normally distributed.
- The population should have equal variances.

- Parametric test require quantitative measurement of sample data in form of an interval or ratio scale, so that arithmetic operations can be used.

2. **Non-parametric tests** are referred to population free or distribution free tests, as they are not based on the characteristics or normality assumptions of the population. The non-parametric tests are applied when the population is not normal or the data being measured is qualitative in nature. These tests can be applied to nominal and ordinal scale data and are easy to compute when the sample size is small. The important non-parametric tests are: the Chi-square test, Runs test, Mann-Whitney test, Wilcoxon matched-pairs signed rank test, Kruskal-Wallis test, Friedman test and Spearman's rank correlation.

9. TEST OF SIGNIFICANCE FOR LARGE SAMPLE Z-TEST

The variables are classified into small and large sample. A sample with more than 30 variables is known as large sample. In case when sample size is large i.e., ($n \geq 30$), z-test can be used for testing the hypothesis. Z-test is a popular test for judging the significance of mean and proportion. Z value will be computed from the sample data with the help of z statistics. Then the critical Z value is found out from the table at a particular level of significance. The Table Z value will be compared with calculated Z statistics for taking the decision about the hypothesis. If the calculated Z value is more than the Table value then reject the null hypothesis and conclude that the alternative hypothesis is accepted or vice-versa.

9.1 Z-test for Single Mean (σ is known)

In this case a single mean is drawn from a particular population and some measure of central tendency like mean is calculated in order to determine the statistical significance between sample mean and population mean.

The z-test is used by applying the following formula:-

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

Where, $\overline{X}$ = The Mean of Sample,

μ = Population Mean.

σ = Standard Deviation of Population.

n = The number of items in a sample.

Example-1. The mean life of time of 225 CFL bulbs produced by the a company is found to be 1270 hours with a standard deviation of 90 hours. Test the hypothesis that it has mean life of 1300 hours.

Solution- Null Hypothesis: $\overline{X} = \mu = 2600$.

Alternative Hypothesis: H₁: $\overline{X} \neq \mu \neq 2600$

$\overline{X} = 1270$, $\mu = 1300$, $\sigma = 150$, $n = 225$.

$$Z = \frac{1300 - 1270}{150 / \sqrt{225}}$$

$$Z = \frac{30}{90/15}$$

$$Z = \frac{30}{6}$$

$$Z = 5$$

Decision: The calculated value 5 is more than standard value 1.96 at 5% level of significance. Hence, hypothesis is rejected and there is a significant difference in sample mean and population mean.

9.2 Z-test for Testing the Difference Between Two Mean

There are many instances in social science and business research which involve a comparison between two populations. The null hypothesis is such situation is stated as under:-

Null Hypothesis: There is no significant difference between mean of two populations.

$$H_0: \overline{X} = \overline{Y}$$

Alternative Hypothesis: There is significant difference between mean of two populations.

$$H_0: \overline{X} \neq \overline{Y}$$

To test the hypothesis the following formula of z-test is applied:-

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\left\{\frac{\sigma_x^2}{n1} + \frac{\sigma_y^2}{n2}\right\}}}$$

Where, $\overline{X}$ = Mean of the first sample.

$\overline{Y}$ = Mean of the second sample

σ_x^2 = Variance of the first population.

σ_y^2 = Variance of the second population.

N_1 = Size of first Sample.

N_2 = Size of second sample.

Example 2: A random sample of 1000employees from Chandigarh shows their mean wages as Rs 47 per week with standard deviation Rs. 28. A random sample of 1500 employees from Trivandrum gives a mean wage of Rs 49 with standard deviation of Rs. 40 per week. Is there a significance difference between the mean wages of two cities?

Solution:

Let us take a null hypothesis that there is no significant difference between mean wages between two cities.

$$H_0: \overline{X} = \overline{Y}$$

$\overline{X} = 47.$ $\overline{Y} = 49,$ $\sigma_1^2 = 28.$ $\sigma_2^2 = 40.$ $N_1 == 1000.$ $N_2 == 1500.$

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\left\{\frac{\sigma_x^2}{n1} + \frac{\sigma_y^2}{n2}\right\}}}$$

$$Z = \frac{47 - 39}{\sqrt{\left\{\frac{(28)^2}{1000} + \frac{(40)^2}{1500}\right\}}} = 1.47$$

Decision: Calculated value is 1.47 which is less than the standard value 1.96 at 5% level of significance. Hence hypothesis is accepted, which shows that there is no significant difference between mean wages in two cities.

9.3 Z-test for Hypothesis Testing of Difference of two Proportions:

When the two samples are drawn from the different population having similar universes in that case the researcher may be interested to find out whether the difference between the proportion of success is significant or not. Here the hypothesis formulated will be that there is no significant difference between the proportions of successes between two samples.

For testing the difference between the two proportions the following procedure will be applied:-

Step- 1: Formulate the null hypothesis: There is no significant difference between the proportions of successes between two samples. $H_0: p_1 = p_2$.

Step- 2: Calculate the value of p and q, where p is pooled estimate of the proportion in the population:

$$P_0 = \frac{n_1 p_1 + n2p2}{n1 + n2}, \text{or} \frac{x_1 + x_2}{n_1 + n_2} \quad q_0 = 1 - p_0.$$

Where, x_1 and x_2 stand for the number of occurrences in the two samples of size n_1 and n_2.

Proportion in the first sample = p_1, Proportion in the second sample = p_2

Step- 3: Calculate the standard error of the difference between proportions;

$$\text{S.E.}_{(p1 - p2)} = \sqrt{p_0 q_0 \left\{\frac{1}{n1} + \frac{1}{n2}\right\}}$$

Step- 4: then apply the rule to calculate the z value:

$$Z = \frac{\text{Difference}}{\text{S.E.}} = \frac{p1 - p2}{\text{S.E.}(p1 - p2)}$$

Step- 5: Interpret the results: If the calculated value of Z is more than the table value at particular level of significance (generally, 1.96 at 5% level of significance), the hypothesis will be rejected and it can be concluded that there is a significant difference between the proportions of successes between two samples or vice-versa.

Example 3: A machine has produced 20 defective items in a batch of 500 items. After repairing the machine, it produced 3 defective items in a batch of 100 items. Find out whether there is any improvement in the machine after the repair.

Solution:

Null hypothesis: there is no significant difference in the production of defective items before and after the repair.

$H_0: p_1 = p_2.$

Step- 2. $$P_1 = \frac{20}{500} = 0.04, \quad P_2 = \frac{3}{100} = 0.03,$$

Step- 3. $$P_0 = \frac{500 \times 0.04 + 100 \times 0.03}{500 + 100} = \frac{20+3}{500+100} = \frac{23}{600} = 0.0383.$$

$$q_0 = 1 - p_0. \quad q_0 = 1 - 0.383 = 0.9617$$

Step- 4: Calculate the standard error of the difference between proportions;

$$\text{S.E.}_{(p1 - p2)} = \sqrt{0.383 \times 0.9617 \left\{ \frac{1}{500} + \frac{1}{100} \right\}} = 0.021.$$

Step- 5. Apply the rule to calculate the z value:

$$Z = \frac{\text{Difference}}{\text{S.E.}} = \frac{p1 - p2}{\text{S.E.}(p1 - p2)} = \frac{0.04 + 0.03}{0.021} = 0.476.$$

Step- 6. Interpretation of results: The calculated value of Z is 0.476 which is less than the table value 1.96 at 5% level of significance, hence, the hypothesis will be accepted and it can be concluded that there is no significant difference in the production of defective items before and after the repair. Machine has not improved significantly.

10. TEST OF SIGNIFICANCE FOR SMALL SAMPLE T-TEST

Small sample size referred to size of sample which is less than 30. In case of small sample size the z-test is not appropriate test statistic as the assumptions on which it is based do not hold good in case of small sample. The theoretical work on t-distribution was done by W.S. Gosset (1876-1937) under the pen name "student' as he was the employee of the company Guinness & Sons, a Dublin bravery, Ireland, which did not allowed it employees to publish research findings under their own names. The t-distribution is used when sample size is less than 30 and the population standard deviation is not known.

't'- test can also be applied to estimate the population mean μ when population standard deviation is unknown and sample size is even large irrespective of shape of population (Naval Bajpai-2011). There is debate on this issue. Some researchers apply the 't'- test when population standard deviation is unknown, irrespective of sample size. Some of the other researchers are of the opinion that for large sample size the 't'- test and 'Z'- test are almost identical even when population standard deviation is unknown (Zikmund, Babin, Carr, Adhikari and Griffin - 2014). When sample sizes approaches 120, the sample standard deviation becomes a very good estimate of the population standard deviation: beyond 120 sample size, t-test and Z-test are virtually identical (Donald, Pamela S Schindler and J K Sharma - 2014).

The t-statistic is defined as under:-

$$t = \frac{\overline{X} - \mu}{s} \sqrt{n},$$

$\overline{X}$ = The mean of sample.

μ = The actual or hypothetical mean of the population.

S = Standard deviation of sample.

n = Sample size.

Degree of Freedom: Degree of freedom is used to see the table value for testing the hypothesis as $V = n-1$. If hypothesis is to be tested 5% level of significance under one tail, then the value is to be seen below the 0.025 level. If the value of table is two tails, then the value is to be seen below the 0.05.

10.1 Assumptions of t-test

The following are the pre-requisites for the application of t-test:-

1. The population from which a sample is drawn is normal.
2. The samples have been drawn at random.
3. The population standard deviation is not known.
4. Sample size should small i.e., less than 30.

10.2 Properties of t-test

Following are the properties of the t-test:-

1. It gives a normal bell shaped curve which is symmetrical.
2. The mean of t-distribution is zero.
3. Its value range between plus (+) infinity to minus (--) infinity.
4. Its variance is greater than one and as the sample size increases it tends to move towards unity.
5. t-test can be used even for large sample but the large sample theory can't be used for small sample.
6. It has larger area at tails than normal distribution, and so, it assigns higher probabilities to extreme outcomes.
7. It has a set of critical values like the distribution of Z, X^2, and F etc. These values are referred to, for testing the magnitude of the calculated value of the test statistics.
8. It is higher than the normal distribution at the both side of the tails of the curve but lower in height at the point of the mean (μ).
9. The constant C is actually a function of V (pronounced as nu), so that for a particular value of V, the distribution of $f(t)$ is completely specified. Thus, $f(t)$ is a family of functions, one for each value of V.

10.3 The application of the t-test

The following are the cases in which t-test is generally used to test the significance of the various results obtained from small sample:-

1. To test the significance of the mean of random sample.
2. To test the difference between means of two samples (Independent Samples).

3. To test the difference between means of two samples (Dependent Samples) or Paired t-test.
4. To test the significance of observed correlation coefficient.

insignificant.

11. F-TEST AND ITS APPLICATION

F-test is another important parametric test in addition to t-test and z-test which is used to find ou whether the two independent estimates of population variance differ significantly or whether the tw samples may be regarded as drawn from normal populations having the same variance. This technique can be used advantageously in the analysis of variances involving two or more number of samples where the z-test and t-test can't be used because these two tests are used in case of two samples only. The credit of developing F-test goes to famous statistician R. A Fisher, so F-test is named in his honour. F-test can be used to test the equality of variance of two normal populations, analyse variance for than two independent samples as well analysis of covariance.

The F-distribution is the distribution of ratio of variances. This test is applied to test the equality of variances of two populations. It is also used for testing the significance of regression equation and testing the equality of several population means.

The procedure of calculating F-test is elaborated as below:-

Null hypothesis: The two populations have same variance or there is no significant difference in the variances of two populations. Ho : $\sigma_1^2 = \sigma_2^2$.

Alternative hypothesis: The two populations have not same variance or there is significant difference in the variances of two populations. Ho : $\sigma_1^2 \neq \sigma_2^2$.

F-ratio is calculated as under:

$$F = \frac{\text{Large estimate of variance}}{\text{Small estimate of variance}}$$

$$F = \frac{S_1^2}{S_2^2}$$

Where, S_1^2 = Greater variance . S_2^2 = Smaller variance.

Degree of freedom: $V_1 = n_1 - 1$ and $V_2 = n_2 - 1$

V_1 = Degree of freedom for sample having larger variance.

V_2 = Degree of freedom for sample having smaller variance.

Decision: The decision whether to accept or reject the null hypothesis is based on the following criteria:

Acceptance criteria: If the calculated value of F-ratio is less than table value of F at particular level of significance, then F- ratio is not significant and null hypothesis is accepted. Hence, the two populations have same variance or there is no significant difference in the variances of two populations.

Rejection criteria: If the calculated value of F-ratio is more than table value of F at particular level of significance, then F- ratio is significant and null hypothesis is accepted. Hence, two populations have not same variance or there is significant difference in the variances of two populations.

Example 8: In a laboratory experiment, two random samples gave the following results:-

Sample	Size	Sample	Sum of Squares of Deviation from the mean
1	10	15	90
2	13	14	108

Test the equality of sample variance at 5% level of significance. The table value for $V_1 = 9$ and $V_2 = 12$ is 2.7964.

Solution:

Null Hypothesis: There is no significant difference in the variance of two samples.

Ho : $\sigma_1^2 = \sigma_2^2$.

Alternative Hypothesis: There is significant difference in the variance of two samples.

Ho : $\sigma_1^2 \neq \sigma_2^2$

By applying the F-Test,

$$F = \frac{S_1^2}{S_2^2},\ S_1^2 = \frac{90}{10-1} = \frac{90}{9} = 10. \quad S_2^2 = \frac{108}{13-1} = \frac{108}{12} = 9.$$

$$F = \frac{10}{9} = 1.11$$

Decision: The table value for $V_1 = 9$ and $V_2 = 12$ is at 5% level of significance is 2.7964. The calculated value—1.11, is less than table vale, hence hypothesis is accepted. It can be concluded that there is no significant difference in the variance of two samples.

12. ANALYSIS OF VARIANCE

The analysis of variance, popularly known as ANOVA is very useful statistical technique for testing the equality of more than two means of populations. As discussed in the previous chapter, the significance of the difference between the two means of two samples can be judged through the application of either Z-test or t-test, but the difficulty arises when the researcher has to test the significance of difference among the more than two sample means at the same time. This technique is successfully used by the researchers in the field of economics, commerce, management, education, psychology, sociology and in many other areas. This technique is used where multiple sample cases are involved.

The ANOVA technique is developed by Professor R. A. Fisher in 1920s and later on Professor Snedecor and many others contributed in the development of this technique. The analysis of variance technique consists of classifying and cross classifying statistical results and testing whether the means of a specified classification differ significantly. In other word, ANOVA is a procedure for testing the difference among different groups of data for homogeneity. ANOVA is a technique that separates the total amount of variation in a set of data in two ways, the amount which can be attributed to chance and the amount which can be attributed to the specific causes. There may be variation between samples and also within sample items. The between sample variance represents the effect of treatment or factor and within sample variance describes the deviation of data points within each group from the sample mean which is often called error. It consists of splitting the variance for analytical purpose. Thus, it is a technique of analysing the variance to which a response is subject into its various components corresponding to various sources of variation. ANOVA is also used to test the significance of regression analysis.

The definitions of analysis of variance are given as under:-

In the words of R. A. Fisher, "The analysis of variance is defined as the separation of the variance ascribable to one group of causes from the variance ascribable to other groups".

In the words of H. T. Hayslett, "The analysis of variance is a technique that separates the variation that is present into independent components, then these components are analysed in order to test certain hypothesis".

Hence, ANNOVA is a test, the purpose of which is to assess the plausibility of the hypothesis stating that the means of normal distributions are indeed equal. The observations in the sample data may be classified according to one factors criterion or two factors criterion. The classification according to one factor criterion is called one-way classification and the classification according to two factor criterion is called two-way classification.

12.1 Assumptions of Analysis of Variance

F-test is based on the following assumptions:

1. Normality: The population is normally distributed.

2. Homogeneity: The variance within each group should be equal for all groups. This assumption ensure that all variances within the group are clubbed into single "within group" source of variation.

3. Randomness: The samples under study have been drawn at random from the population.

4. Independence of Error: This implies that variation of each item around the group mean should be independent for each item value.

5. Positive F Value: Since F-distribution is always a ratio of square value, so it can never be of a negative value.

6. Equal Variance: It begins with null hypothesis that variance or means of all populations are equal.

CHAPTER

14

Chi-Square Test and Factor Analysis

Learning Objective

After the completion of this chapter students should be able to understand the meaning of Chi-square test and Factor Analysis. In addition to it some practical problems have also been discussed to make the application of concept clear.

1. INTRODUCTION

The various tests of significance like Z-test, t-test and F-test are based on certain assumptions or parameters that sample were drawn from normally distribute populations. However, there are many situations in which it is not possible to make any assumption about the distribution of the population from which samples are drawn. In such circumstances the non-parametric tests such as Chi-square test, Sign test, Rank sum test, Rank correlation, Median test, One sample run test, H-test, Willcoxon paired comparison test and Kolmoggrov Semirrnor are widely used in behavioural science that do not require any assumption of normalcy about the parameters or the population values.Generally, the most widely non-parametric test used is Chi-square test of independence and goodness of fit.

The Chi-square test is applied to evaluate the significance of difference between a set of the observed frequencies and a set of the corresponding expected frequencies of a sample drawn without any assumption about its parent distribution. It is denoted by the Greek letter X^2, which is read as Chi-square. This statistical technique was developed by Professor Karl Pearson of England in the year 1900, and was used initially in sociological and psychological researches. Like other tests of significance the Chi-square test is also function of degree of freedom, as the number of degree of freedom increases the Chi-square distribution becomes more symmetric. Chi-square test can never have a negative value being a sum of square quantities. The value of X^2 will be between zero (0) to infinity (∞).

The results of X^2 describe the magnitude of the discrepancy between theory and observation, which is defined mathematically as below:-

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

Where, O = Observed Frequencies.

E = Expected Frequencies.

Hence, X^2 test may refers to a "non-parametric test statistics which measure the significance of difference between a set of the observed frequencies and their corresponding expected or theoretical frequencies relating to a problem that does not require any assumption of normalcy about the parameters, or parent distribution".

2. CHARACTERISTIC OF CHI-SQUARE TEST

The essential characteristics of Chi-square test that emerges out of the above mentioned discussion are cited as under:-

1. It is non-parametric test statistics or distribution free test.
2. It is not symmetric.
3. The value of Chi-square test is always positive being a sum of square quantities. Thus the value of X^2 ranges between zero (0) to infinity (∞). If the value of X^2 is zero it means the observed and expected frequencies completely coincide. The greater the value of X^2 the greater will be the difference between the observed and expected frequencies.
4. It evaluates the significance of difference between a set of the observed frequencies and their corresponding expected or theoretical frequencies by making the comparison between its calculated value and Table value with reference to degree of freedom at desired level of significance and thereby enables the researcher to draw a conclusion whether difference is significant or insignificant.
5. The shape of Chi-square depends up on the degree of freedom degree of freedom, as the number of degree of freedom increases the Chi-square distribution becomes more symmetric or less skewed and this has been presented in the following figure:-

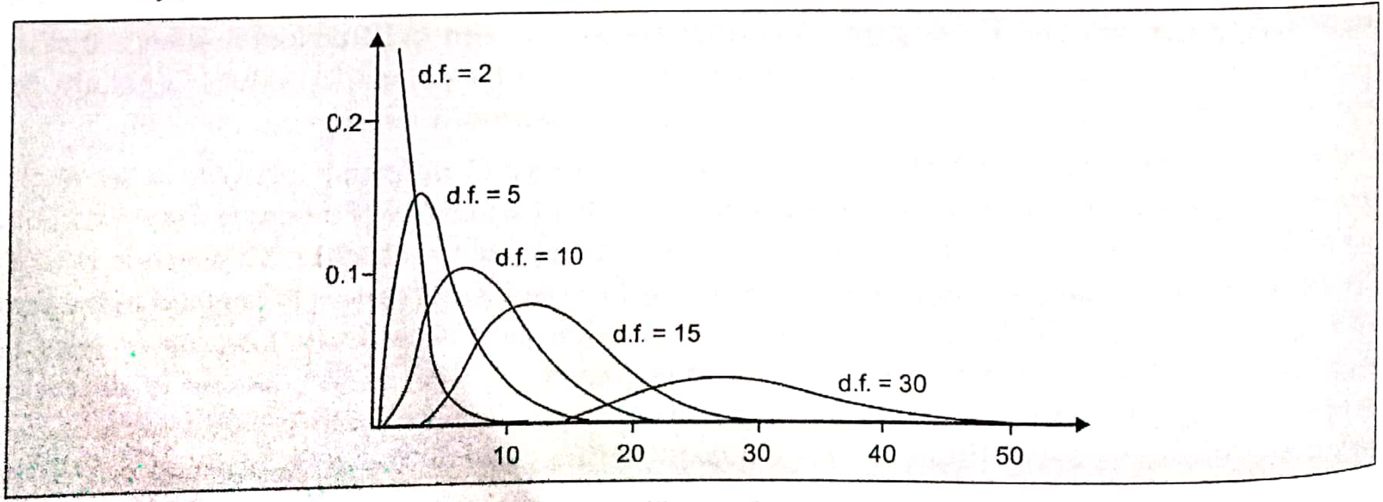

Figure – 1

The above mentioned figure shows that chi-square is becoming more symmetric with the increase in the degree of freedom. When the degree of freedom was 2 then the chi-square was skewed and when the degree of freedom increased to 30 then the chi-square become more symmetric.

3. ASSUMPTION OF CHI-SQUARE TEST

The following conditions must be satisfied before applying the X^2 test:-

1. The total of frequencies of the given distribution is fairly large i.e., N ℘ 50.
2. No frequency cell is less than 5. In case any cell is less than 5 it is to be grouped with lower or upper cell by the technique of pooling to make it 5 or more or preferably more than 10.
3. The observations are independents. This implies that no individual item is included twice or more in the sample.
4. Data should be in original units and are not replaced by any relative frequencies like proportions, percentages, ratio, rates etc.
5. No assumption of normalcy is there about the population values or their parameters. It is distribution free statistical test and applied when data is of qualitative nature.
6. The constraints on cell frequencies, if any, are linear.
7. If parametric test like Z-test, t-test and F-test are applied to the problem then X^2 test should not be applied.
8. The values of independent and dependent variables must be mutually exclusive.
9. The total of expected frequencies is equal to total of the observed frequencies.
10. The outcomes are independent of each other.

4. PROCEDURE OF CHI-SQUARE TEST ANALYSIS

The following procedure is adopted in the X^2 test analysis:-

1. Set up the null hypothesis that there is no significant difference between the observed and expected values.

   H_0: There is no significant difference between the observed and expected values or the opinion of respondents is equally distributed.

   Alternative Hypothesis:

   H_1: There is significant difference between the observed and expected values or the opinion of respondents is not equally distributed.

2. Set the level of significance.
3. Arrange the observed values or frequencies in a column and denote them by O.
4. Calculate the expected frequencies -- E values in each category by multiplying the category probability by sample size.

$$E = \frac{RT \times CT}{N}$$

Where, E = Expected Frequencies.

RT = The row total for the row containing the cell.

CT = The column total for the column containing the cell.

N = The total number of observations

5. Calculate the difference between observed and expected frequencies and obtain the squares of these differences i.e., **$(O - E)^2$.**

6. Divide the each squared difference by the respective expected frequency E and obtain the total as $\frac{(O \times E)2}{E}$. This gives the value of X^2.

7. Obtain the critical value of X^2 from the X^2 table with reference to the degree of freedom and the desired level of significance.

8. Compare the calculated value and table value. If the calculated value of X^2 is more than the table value of X^2 the null hypothesis will be rejected and it can be concluded that there is significant difference between the observed and expected values and this difference is not due to fluctuations of sampling. If the calculated value of X^2 is less than the table value of X^2 the null hypothesis will be accepted and it can be concluded that there is no significant difference between the observed and expected values and the difference is due to fluctuations of sampling.

5. USES OF X^2 TEST

The uses of X^2 test can be elaborated as under:-

5.1 X^2 Test for Independence of the Attributes

Chi-square test of independence is used to determine the relationship or association between two or more than two variables. In a business a market researcher might be interested in knowing the relationship between two variables or check whether they are independent of each other or not. For example, a cosmetic company may be interested in knowing whether the purchase of cosmetic is independent of the customer's age or whether it is dependent on the customer's age. X^2 test can also be used in checking the effectiveness of any medicine or association between inter-cultural marriages and divorce or association between the intelligence and gender etc.

Example -1: From the following data of NHPC Limited use X^2 – test and conclude whether Yoga as stress management technique is effective in managing stress or not.

Stress Management Technique	Affected	Not-affected	Total
Used	140	30	170
Not Used	60	20	80
	200	50	250

Solution:

Null Hypothesis: Yoga as stress management technique is not effective in managing stress.

Expected frequency of Cell A = $E = \frac{RT \times CT}{N} = E = \frac{200 \times 170}{250} = 136.$

Expected frequency of Cell B = $E = \frac{RT \times CT}{N} = E = \frac{200 \times 80}{250} = 64.$

Expected frequency of Cell C = $E = \frac{RT \times CT}{N} = E = \frac{50 \times 170}{250} = 34.$

Expected frequency of Cell D = E = $\frac{RT \times CT}{N} = E = \frac{50 \times 80}{250} = 16.$

Determination of Calculated value of the X^2

Observed frequency (O)	Expected frequency (O)	Difference (O–E)	Square of differences $(O–E)^2$	Fraction of differences $\frac{(O \times E)2}{E}$.
140	136	4	16	0.118
60	64	-4	16	0.250
30	34	4	16	0.471
20	16	-4	16	1.000
ΣO = 168	ΣE= 168	0		1.839

$$X^2 = \sum \frac{(O-E)2}{E} = 1.839$$

The degree of freedom is v = (r-1) (c-1) = (2-1)(2-1).= 1 Level of significance is 5%.

The Table value of X^2 is for v= 1 and p = 0.05 is 3.84.

Decision: The calculated value of X^2 is 1.839 which is less than of its Table value i.e., 3.84, the null hypothesis is accepted and it can be concluded that Yoga as stress management technique is not effective in managing stress.

7. FACTOR ANALYSIS

Factor analysis is a widely used multivariate interdependence technique in the field of commerce, management, economics, social and behavioural sciences. This technique is applicable when there is a systematic interdependence among a set of observed variables and the investigator is interested in knowing something more fundamental or latent which creates commonality. It is termed as data reduction and summarization technique which reduce a large number of variables to the more manageable variables. Data reduction, from a large set of variables is based on the nature and character of relationship among the variables. For instance, in marketing research, there may be a large number of variables, most of which are correlated and that must be reduced to a manageable level. The relationship among set of variables are systematically examined on the basis of commonality among these variables and presented in the form of a few underlying factors.

In the factor analysis the variables are grouped into a number of factors based on the degree of correlation among the variables and the factor so derived may be treated as new factors which are termed as latent variable. The meaning and name of such variable is subjectively determined by the researcher. Hence, factor analysis is a statistical technique used to determine the prescribed number of uncorrelated factors, where each factor is obtained from a list of correlated variables. This analysis may be applied to examine the underlying patterns or relationship for a large number of variables and to determine whether the data can be condensed or summarized into a smaller set of factors.

The factor analysis can be performed mainly for two reasons:-

1. To indentify a new, smaller set of uncorrelated variables to be used in subsequent multiple regression analysis.
2. To identify underlying factors that are unobservable but explain correlation among a set of variables.

8. ESSENTIAL FEATURES OF FACTOR ANALYSIS

The under mentioned are some characteristics of factor analysis:-

1. It is multivariate technique where initially large numbers of variables are considered.
2. The large numbers of variables are reduced or summarized to a small number of factors on the basis of some commonality.
3. Selection of variables constituting a factor is made on the basis of total variation accounted for by them.

4. It also provides a measure of correlation. A large number of responses about a large number of variables are used to identify the highly correlated variables which constitute a factor.
5. The variables constituting a factor are highly correlated with one another in a factor but are not correlated with other factor.
6. For each factor the researcher has to exercise his judgement to determine the name or theme of the factor.
7. Factor models are based on the assumption of linear relationship.
8. While deciding on the number of variables to be included in factor analysis, researchers need to ensure that a group of minimum 4 to 5 variables forms one factor. Hence, the factor analysis must have good number of variables, depending up on the study also.
9. The decision about the sample size is also important in this context. However, the minimum sample size has to be more than 100. Some researchers propose that average sample size for factor analysis is to be at least 10 times of number of variables. Whereas in some studies average sample size of 20 or more has been suggested by some researchers.

9. ASSUMPTIONS OF FACTOR ANALYSIS

Following are the assumptions of factor analysis:-

1. The factors are derived so as to minimise the correlations among the large number of variables. Hence, the highly correlated variables are combined within a single factor.
2. The variables included in factor analysis can construct one or more latent factor.
3. The variables are interrelated, however does not assume any causality.
4. The sample is representative of population. Sample is homogeneous and shows diversity in the characteristics across all respondents uniformly.
5. Factors can be represented as a linear combination of the variables.
6. The mean of unique factor is zero. They are known as error and these errors are random.
7. The correlation between indicator variables can be attributed to common factors.
8. The common factors are standardized. This means that mean of each factor is zero and standard deviation is one.
9. Each specific factor has its unique variance.
10. The variables allotted to a factor are nearly independent of the variables allocated to other factors.
11. The specific factors are uncorrelated with the common factors.
12. The factors are derived in such a manner that the percentage of total variance explained by the variables and included in successive factors is maximised.

10. APPLICATION OF FACTOR ANALYSIS

Factor analysis is used in many situations in business research. The uses and application of factor analysis is summed up as under:-

1. **Reduction in the Number of Variables or Parsimony:** Factor analysis reduces the number of variables taken initially on the basis of some common characteristics or nature to a small number of independent factors. Multi-collinearity is generally preferred between the variables as the correlations are the key to the data reduction. It is also called as data reduction and summarization technique.
2. **Determination of Latent Factors:** Factor analysis identifies latent factors which determine the correlation among a set of variables.
3. **Determination of Grouping of Factors:** Factor analysis identifies the latent relationship among the various groups of variables under observation. Sometimes a certain relationship exists among the variables but it is not clearly observable but it can be highlighted by the technique of factor analysis.
4. **Helps in Subsequent Analysis:**Factor analysis identifies a new, smaller set of uncorrelated variables in the place of original larger set of correlated variables for subsequent multivariate analysis.
5. **Determination of Cluster of Observation:** Factor analysis can be used for determining the cluster among the observation based on a survey.
6. **Developing Composite Indicators:** Factor analysis can be used for developing composite indicators for the comparison of two situations or two periods, which are considered as very significant in business and economic research.
7. **Scaling:** Factor analysis can also be used in the development of scale in behavioural research. It divides the characteristics into some independent construct or concepts which represents a scale to measure an underlying behavioural dimension.
8. **Market Research:** In the field of marketing research factor analysis helps in identifying the underlying dimensions from the observed variables which helps the researcher to group consumers based on their homogeneity in the characteristics that are measured. This knowledge can be useful in improving the quality of the product, market segmentation, product research, pricing studies, planning for more effective advertisement campaign, control of quality etc.

11. KEY STATISTICS ASSOCIATED WITH FACTOR ANALYSIS

The key statistics associated with factor analysis are explained as under:-

1. **Factor:** Factor is a weighted linear combination of the original variables under study. The output of factor analysis is in terms of factor. It represents the underlying constructs or concept that summarizes or account for the original set of observed variables.
2. **Exploratory Factor Analysis:** This technique is applied when a researcher has no prior knowledge about the factors that the variables will be indicating. In such case, computer based techniques are used to indicate appropriate number of factors.
3. **Confirmatory Factor Analysis:** This technique is applied when a researcher has the prior knowledge about the number of factors the variables will be indicating.
4. **Factor Loading:** Factor loading are the values which explain the extent of closeness of relationship among variables constituting a factor. These are simple variable correlation between the variable and called as factor-variable correlations.

5. **Factor Scores:** Factor scores are the coefficient of the factors.

6. **Factor Matrix:** A factor matrix contains the factor loadings of all the variables on all the factors extracted.

7. **Factor Plot:** This is a plot where the factors are on different axis and the variables are drawn on these axes. This plot can be interpreted only if the numbers of factors are 3 or less.

8. **Factor Rotation:**Rotation of factor is done to reveal different structure of data. The most commonly used technique for rotation is the varimax procedure. Though the different rotation gives different results and should be taken as equal, not superior or inferior to others. For final interpretation the correct rotations should be identified and selected. In case factors are independent, orthogonal rotation is done, but if the factors are correlated then an oblique rotation is done. On rotation commonality remains unaffected but the Eigen value changes.

9. **Eigen Value:** Eigen value for each factor is the total variance explained by each factor.

10. **Scree Plot:**A scree plot is a plot of the Eigen values against the number of factors in order of extraction.

11. **Total Sum of Square:**When the Eigen values of all factors are totalled, the resulting value is termed as total sum of squares.

12. **Communality:** It indicates the amount of variance an original variable shares with all other variables included in the analysis. A relatively high value of communality indicates that a variable has much in common with the other variables taken as a group whereas a low commonality means that the variable does not have a strong relationship with other variables. The most common type of factor rotation is a process called varimax.

13. **Percentage of Variance:** This is the percentage of total variance attributed to each factor.

14. **Kaiser-Meyer-Olkin (KMO) Measure of Sample Adequacy:** This is an index used to examine the appropriateness of factor analysis. This statistics compares the magnitude of the observed correlation coefficient with the magnitude of the partial correlation coefficient. High value of this statistics (Between 0.5to1.0)indicates the appropriateness of factor analysis whereas low value of this statistics (below 0.5) indicates the inappropriateness of factor analysis. Kaiser has presented the range the range as follow: Statistics > 0.9 is marvellous, statistics > 0.8 meritorious, Statistics > 0.7 middling, Statistics > 0.6 mediocre, Statistics > 0.5 miserable and Statistics < 0.5 unacceptable.

15. **Bartlett's Test of Sphericity:** This is the test statistics used to examine the null hypothesis that there is no correlation between the variables or variables are uncorrelated in the population. In other words,this test the null hypothesis whether the population correlation matrix is an identity matrix. It is pertinent to mention here that the presence of identity matrix puts the correctness of factor analysis under suspicion. Using the level of significance generally 5%, the degree of relationship among the variables can be identified. A value below 0.05 rejects the null hypothesis and indicates that the data in hand do not produce an identity matrix. This means that there exists a significant relationship among the variables, taken from the factor analysis.