Course Code: ECOPG-402
Foundation Course in Research Methods & Data Analysis-II

DATA ANALYSIS AND INTERPRETATION

1. What is hypothesis?

Ordinarily, when one talks about hypothesis, one simply means a mere assumption or some supposition to be proved or disproved. But for a researcher hypothesis is a formal question that he intends to resolve. Thus a hypothesis may be defined as a proposition or a set of proposition set forth as an explanation for the occurrence of some specified group of phenomena 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 variable. For example, consider statements like the following ones:
"Students who receive counselling will show a greater increase in creativity than students not receiving counselling" Or "the automobile A is performing as well as automobile B."
These are hypotheses capable of being objectively verified and tested. Thus, we may conclude that a hypothesis states what we are looking for and it is a proposition which can be put to a test to determine its validity.

2. Discuss the Characteristics of hypothesis in Research Methodology

Characteristics of hypothesis: Hypothesis must possess the following characteristics:

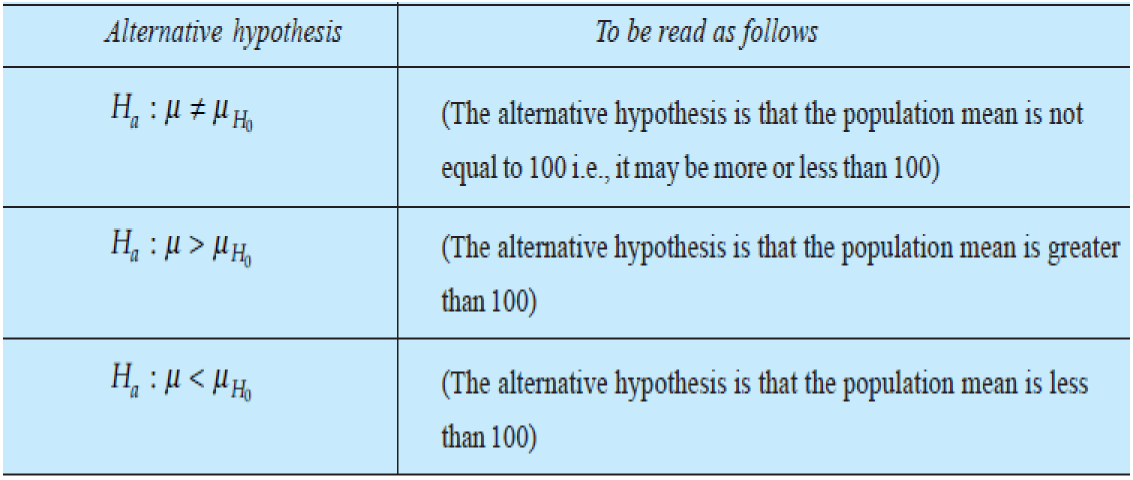
a. Hypothesis should be clear and precise. If the hypothesis is not clear and precise, the inferences drawn on its basis cannot be taken as reliable.
b. Hypothesis should be capable of being tested. In a swamp of untestable hypotheses, many a time the research programmes have bogged down. Some prior study may be done by researcher in order to make hypothesis a testable one. A hypothesis "is testable if other deductions can be made from it which, in turn, can be confirmed or disproved by observation."
c. Hypothesis should state relationship between variables, if it happens to be a relational hypothesis.
d. Hypothesis should be limited in scope and must be specific. A researcher must remember that narrower hypotheses are generally more testable and he should develop such hypotheses.
e. Hypothesis should be stated as far as possible in most simple terms so that the same is easily understandable by all concerned. But one must remember that simplicity of hypothesis has nothing to do with its significance.
f. Hypothesis should be consistent with most known facts i.e., it must be consistent with a substantial body of established facts. In other words, it should be one which judges accept as being the most likely.
g. Hypothesis should be amenable to testing within a reasonable time. One should not use even an excellent hypothesis, if the same cannot be tested in reasonable time for one cannot spend a life-time collecting data to test it.
h. Hypothesis must explain the facts that gave rise to the need for explanation. This means that by using the hypothesis plus other known and accepted generalizations, one should be able to deduce the original problem condition. Thus hypothesis must actually explain what it claims to explain; it should have empirical reference.

3. Discuss the Basic concepts of Hypothesis testing.

Basic concepts in the context of testing of hypotheses need to be explained.

Null hypothesis and alternative hypothesis: In the context of statistical analysis, we often talk about null hypothesis and alternative hypothesis. If we are to compare method A with method B about its superiority and if we proceed on the assumption that both methods are equally good, then this assumption is termed as the null hypothesis. As against this, we may think that the method A is superior or the method B is inferior, we are then stating what is termed as alternative hypothesis. The null hypothesis is generally symbolized as H0 and the alternative hypothesis as Ha. Suppose we want to test the hypothesis that the population mean bmg is equal to the hypothesised mean mH0 d i = 100. Then we would say that the null hypothesis is that the population mean is equal to the hypothesized mean 100 and symbolically we can express as: $H_0 : \mu = \mu_{H_0} = 100$

If our sample results do not support this null hypothesis, we should conclude that something else is true. What we conclude rejecting the null hypothesis is known as alternative hypothesis. In other words, the set of alternatives to the null hypothesis is referred to as the alternative hypothesis. If we accept H0, then we are rejecting Ha and if we reject H0, then we are accepting Ha. For H0 : m m = = H0 100 , we may consider three possible alternative hypotheses as follows:



Alternative hypothesis	*To be read as follows*
$H_a : \mu \neq \mu_{H_0}$	(The alternative hypothesis is that the population mean is not equal to 100 i.e., it may be more or less than 100)
$H_a : \mu > \mu_{H_0}$	(The alternative hypothesis is that the population mean is greater than 100)
$H_a : \mu < \mu_{H_0}$	(The alternative hypothesis is that the population mean is less than 100)

The null hypothesis and the alternative hypothesis are chosen before the sample is drawn (the researcher must avoid the error of deriving hypotheses from the data that he collects and then testing the hypotheses from the same data). In the choice of null hypothesis, the following considerations are usually kept in view:

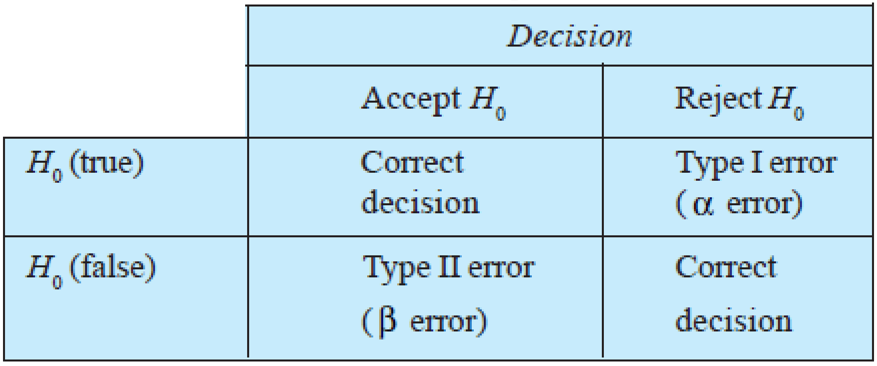
- Alternative hypothesis is usually the one which one wishes to prove and the null hypothesis is the one which one wishes to disprove. Thus, a null hypothesis represents the hypothesis we are trying to reject, and alternative hypothesis represents all other possibilities.
- If the rejection of a certain hypothesis when it is actually true involves great risk, it is taken as null hypothesis because then the probability of rejecting it when it is true is a (the level of significance) which is chosen very small.
- Null hypothesis should always be specific hypothesis i.e., it should not state about or approximately a certain value. Generally, in hypothesis testing we proceed on the basis of null hypothesis, keeping the alternative hypothesis in view. Why so? The answer is that on the assumption that null hypothesis is true, one

can assign the probabilities to different possible sample results, but this cannot be done if we proceed with the alternative hypothesis. Hence the use of null hypothesis (at times also known as statistical hypothesis) is quite frequent.

The level of significance: This is a very important concept in the context of hypothesis testing. It is always some percentage (usually 5%) which should be chosen with great care, thought and reason. In case we take the significance level at 5 per cent, then this implies that H0 will be rejected when the sampling result (i.e., observed evidence) has a less than 0.05 probability of occurring if H0 is true. In other words, the 5 per cent level of significance means that researcher is willing to take as much as a 5 per cent risk of rejecting the null hypothesis when it (H0) happens to be true. Thus the significance level is the maximum value of the probability of rejecting H0 when it is true and is usually determined in advance before testing the hypothesis.

Decision rule or test of hypothesis: Given a hypothesis H0 and an alternative hypothesis Ha, we make a rule which is known as decision rule according to which we accept H0 (i.e., reject Ha) or reject H0 (i.e., accept Ha). For instance, if (H0 is that a certain lot is good (there are very few defective items in it) against Ha) that the lot is not good (there are too many defective items in it), then we must decide the number of items to be tested and the criterion for accepting or rejecting the hypothesis. We might test 10 items in the lot and plan our decision saying that if there are none or only 1 defective item among the 10, we will accept H0 otherwise we will reject H0 (or accept Ha). This sort of basis is known as decision rule.

Type I and Type II errors: In the context of testing of hypotheses, there are basically two types of errors we can make. We may reject H0 when H0 is true and we may accept H0 when in fact H0 is not true. The former is known as Type I error and the latter as Type II error. In other words, Type I error means rejection of hypothesis which should have been accepted and Type II error means accepting the hypothesis which should have been rejected. Type I error is denoted by a (alpha)known as a error, also called the level of significance of test; and Type II error is denoted by b (beta) known as b error. In a tabular form the said two errors can be presented as follows:

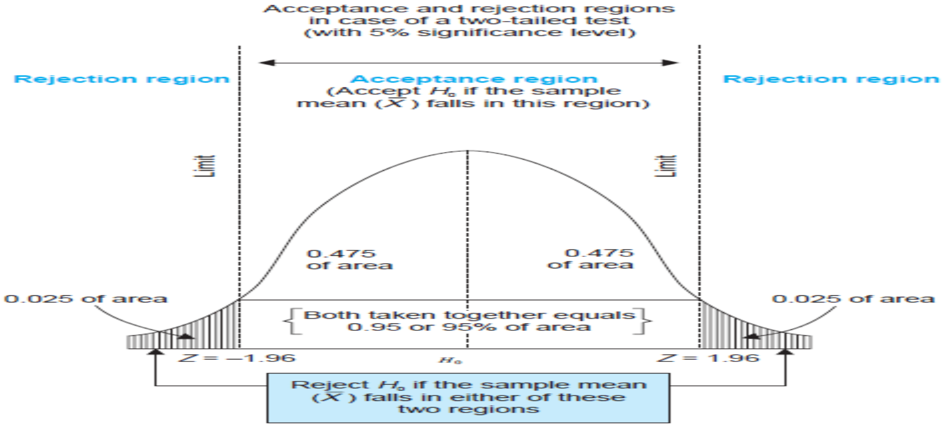


	Decision	
	Accept H_0	Reject H_0
H_0 (true)	Correct decision	Type I error (α error)
H_0 (false)	Type II error (β error)	Correct decision

The probability of Type I error is usually determined in advance and is understood as the level of significance of testing the hypothesis. If type I error is fixed at 5 per cent, it means that there are about 5 chances in 100 that we will reject H0 when H0 is true. We can control Type I error just by fixing it at a lower level. For instance, if we fix it at 1 per cent, we will say that the maximum probability of committing Type I error would only be 0.01.

But with a fixed sample size, n, when we try to reduce Type I error, the probability of committing Type II error increases. Both types of errors cannot be reduced simultaneously. There is a trade-off between two types of errors which means that the probability of making one type of error can only be reduced if we are willing to increase the probability of making the other type of error. To deal with this trade-off in business situations, decision-makers decide the appropriate level of Type I error by examining the costs or penalties attached to both types of errors. If Type I error involves the time and trouble of reworking a batch of chemicals that should have been accepted, whereas Type II error means taking a chance that an entire group of users of this chemical compound will be poisoned, then in such a situation one should prefer a Type I error to a Type II error. As a result one must set very high level for Type I error in one's testing technique of a given hypothesis.2 Hence, in the testing of hypothesis, one must make all possible effort to strike an adequate balance between Type I and Type II errors.

Two-tailed and One-tailed tests: In the context of hypothesis testing, these two terms are quite important and must be clearly understood. A two-tailed test rejects the null hypothesis if, say, the sample mean is significantly higher or lower than the hypothesised value of the mean of the population. Such a test is appropriate when the null hypothesis is some specified value and the alternative hypothesis is a value not equal to the specified value of the null hypothesis. Symbolically, the two tailed test is appropriate when we havewhich may mean m > mH0 or m < mH0. Thus, in a two-tailed test, there are two rejection regions*, one on each tail of the curve which can be illustrated as under:


Acceptance and rejection regions in case of a two-tailed test (with 5% significance level)
Rejection region
Acceptance region
(Accept H_0 if the sample mean ($\bar{X}$) falls in this region)
Rejection region
Limit
Limit
0.475 of area
0.475 of area
0.025 of area
0.025 of area
Both taken together equals 0.95 or 95% of area
$Z = -1.96$
H_0
$Z = 1.96$
Reject H_0 if the sample mean ($\bar{X}$) falls in either of these two regions

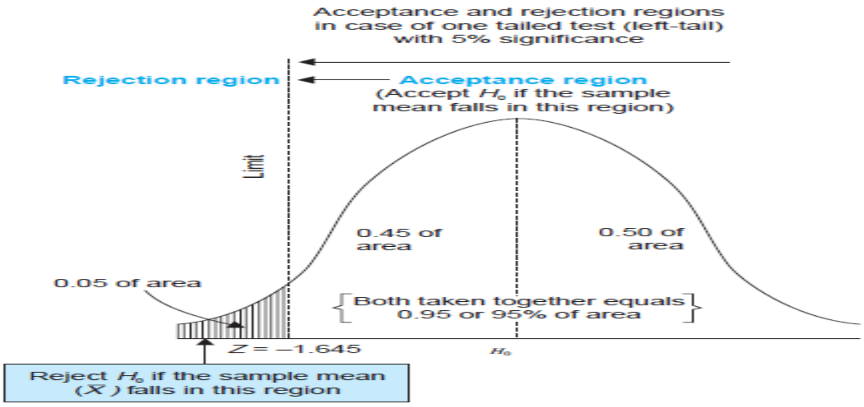
Mathematically we can state:

Acceptance Region A : Z < 1.96

Rejection Region R : Z > 1.96

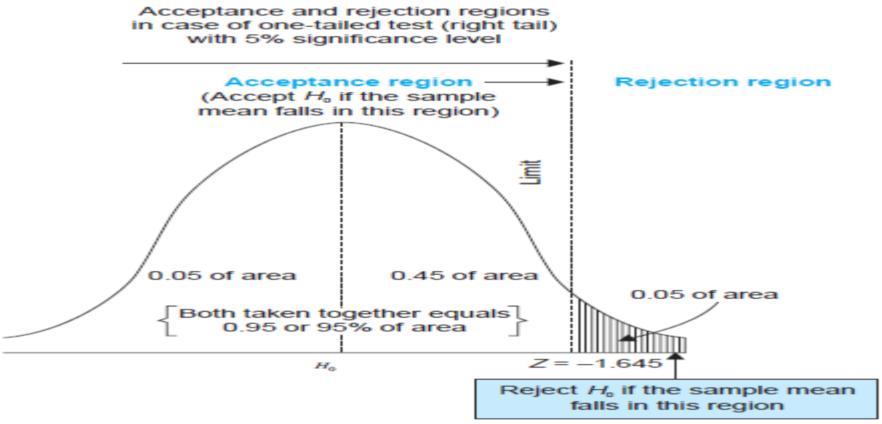
If the significance level is 5 per cent and the two-tailed test is to be applied, the probability of the rejection area will be 0.05 (equally splitted on both tails of the curve as 0.025) and that of the acceptance region will be 0.95 as shown in the above curve. If we take m = 100 and if our sample mean deviates significantly from 100 in either direction, then we shall reject the null hypothesis; but if the sample mean does not deviate significantly from m , in that case we shall accept the null hypothesis.

But there are situations when only one-tailed test is considered appropriate. A one-tailed test would be used when we are to test, say, whether the population mean is either lower than or higher than some hypothesised value. For instance, if our H 0 H0: m = m and Ha H : m < m 0 , then we are interested in what is known as left-tailed test (wherein there is one rejection region only on the left tail) which can be illustrated as below:


Acceptance and rejection regions
in case of one tailed test (left-tail)
with 5% significance
Rejection region
Acceptance region
(Accept H_0 if the sample
mean falls in this region)
Limit
0.45 of
area
0.50 of
area
0.05 of area
Both taken together equals
0.95 or 95% of area
$Z = -1.645$
H_0
Reject H_0 if the sample mean
($\bar{X}$) falls in this region

If our m = 100 and if our sample mean deviates significantly from100 in the lower direction, we shall reject H0, otherwise we shall accept H0 at a certain level of significance. If the significance level in the given case is kept at 5%, then the rejection region will be equal to 0.05 of area in the left tail as has been shown in the above curve.

In case our H0 H0 : m = m and Ha H : m > m 0 , we are then interested in what is known as one tailed test (right tail) and the rejection region will be on the right tail of the curve as shown below:


Acceptance and rejection regions
in case of one-tailed test (right tail)
with 5% significance level
Acceptance region
(Accept H_0 if the sample
mean falls in this region)
Rejection region
Limit
0.05 of area
0.45 of area
0.05 of area
Both taken together equals
0.95 or 95% of area
H_0
$Z = -1.645$
Reject H_0 if the sample mean
falls in this region

4. Evaluate the procedure for hypothesis testing in research methodology in social science research.

Procedure for hypothesis testing means to tell (on the basis of the data the researcher has collected) whether or not the hypothesis seems to be valid. In hypothesis testing the main question is: whether to accept the null hypothesis or not to accept the null hypothesis? Procedure for hypothesis testing refers to all those steps that we undertake for making a choice between the two actions i.e., rejection and acceptance of a null hypothesis. The various steps involved in hypothesis testing are stated below:

1. **Making a formal statement:** The step consists in making a formal statement of the null hypothesis (H_0) and also of the alternative hypothesis (H_a). This means that hypotheses should be clearly stated, considering the nature of the research problem. For instance, Mr. Mohan of the Civil Engineering Department wants to test the load bearing capacity of an old bridge which must be more than 10 tons, in that case he can state his hypotheses. Take another example. The average score in an aptitude test administered at the national level is 80. To evaluate a state's education system, the average score of 100 of the state's students selected on random basis was 75. The state wants to know if there is a significant difference between the local scores and the national scores.
The formulation of hypotheses is an important step which must be accomplished with due care in accordance with the object and nature of the problem under consideration. It also indicates whether we should use a one-tailed test or a two-tailed test. If Ha is of the type greater than (or of the type

lesser than), we use a one-tailed test, but when Ha is of the type “whether greater or smaller” then we use a two-tailed test.

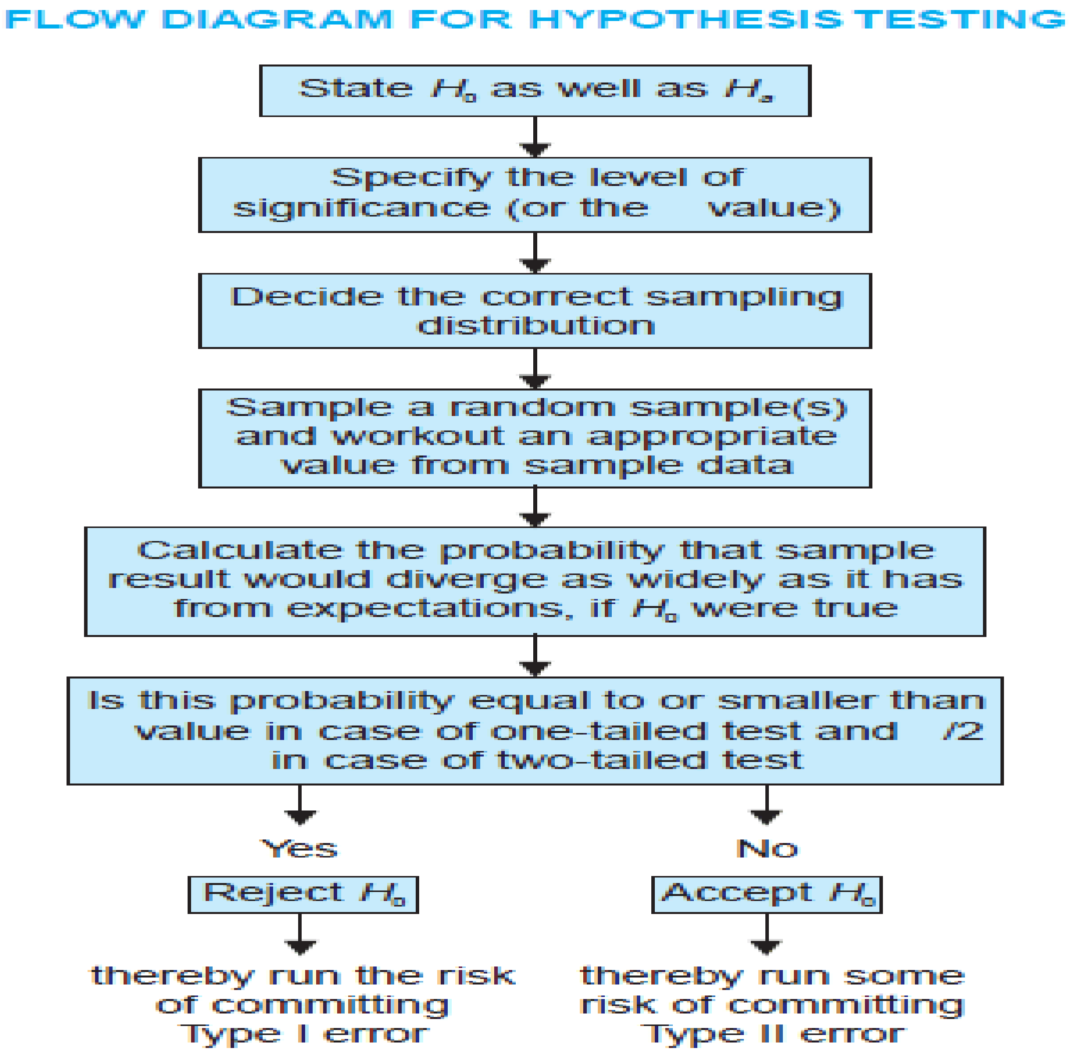
2. **Selecting a significance level:** The hypotheses are tested on a pre-determined level of significance and as such the same should be specified. Generally, in practice, either 5% level or 1% level is adopted for the purpose. The factors that affect the level of significance are: the magnitude of the difference between sample means, the size of the samples, the variability of measurements within samples; and whether the hypothesis is directional or non-directional (A directional hypothesis is one which predicts the direction of the difference between, say, means). In brief, the level of significance must be adequate in the context of the purpose and nature of enquiry.
3. **Deciding the distribution to use:** After deciding the level of significance, the next step in hypothesis testing is to determine the appropriate sampling distribution. The choice generally remains between normal distribution and the t-distribution. The rules for selecting the correct distribution are similar to those which we have stated earlier in the context of estimation.
4. **Selecting a random sample and computing an appropriate value:** Another step is to select a random sample(s) and compute an appropriate value from the sample data concerning the test statistic utilizing the relevant distribution. In other words, draw a sample to furnish empirical data.
5. **Calculation of the probability:** One has then to calculate the probability that the sample result would diverge as widely as it has from expectations, if the null hypothesis were in fact true.
6. **Comparing the probability:** Yet another step consists in comparing the probability thus calculated with the specified value for a , the significance level. If the calculated probability is equal to or smaller than the a value in case of one-tailed test (and a /2 in case of two-tailed test), then reject the null hypothesis (i.e., accept the alternative hypothesis), but if the calculated probability is greater, then accept the null hypothesis. In case we reject H0, we run a risk of (at most the level of significance) committing an error of Type I, but if we accept H0, then we run some risk (the size of which cannot be specified as long as the H0 happens to be vague rather than specific) of committing an error of Type II.

5. Discuss hypothesis testing flow chart diagram.

The above stated general procedure for hypothesis testing can also be depicted in the from of a flowchart for better understanding as shown in Fig. below:

Hypothesis Testing Flow Chart Diagram

Flow Chart diagram for Hypothesis testing is given below


FLOW DIAGRAM FOR HYPOTHESIS TESTING
State H_0 as well as H_a
Specify the level of significance (or the value)
Decide the correct sampling distribution
Sample a random sample(s) and workout an appropriate value from sample data
Calculate the probability that sample result would diverge as widely as it has from expectations, if H_0 were true
Is this probability equal to or smaller than value in case of one-tailed test and /2 in case of two-tailed test
Yes
No
Reject H_0
Accept H_0
thereby run the risk of committing Type I error
thereby run some risk of committing Type II error

6. Short note on Testing of hypothesis

As has been stated above that hypothesis testing determines the validity of the assumption (technically described as null hypothesis) with a view to choose between two conflicting hypotheses about the value of a population parameter. Hypothesis testing helps to decide on the basis of a sample data, whether a hypothesis about the population is likely to be true or false. Statisticians have developed several tests of hypotheses (also known as the tests of significance) for the purpose of testing of hypotheses which can be classified as:

a. Parametric tests or standard tests of hypotheses; and
b. Non-parametric tests or distribution-free test of hypothesis.

Parametric tests usually assume certain properties of the parent population from which we draw samples. Assumptions like observations come from a normal population, sample size is large, assumptions about the population parameters like mean, variance, etc., must hold good before parametric tests can be used. But there are situations when the researcher cannot or does not want to make such assumptions. In such situations we use statistical methods for testing hypotheses which are called non-parametric tests because such tests do not depend on any assumption about the parameters of the parent population. Besides, most non-parametric tests assume only nominal or ordinal data, whereas parametric tests require measurement equivalent to at least an interval scale. As a result, non-parametric tests need more observations than parametric tests to achieve the same size of Type I and Type II errors.4 We take up in the present chapter some of the important parametric tests, whereas non-parametric tests will be dealt with in a separate chapter later in the book.

7. What is the Power of a Hypothesis test?

The power of Hypothesis test is the probability of rejecting null hypothesis .As stated above we may commit Type I and Type II errors while testing a hypothesis. The probability of Type I error is denoted as a (the significance level of the test) and the probability of Type II error is referred to as b . Usually the significance level of a test is assigned in advance and once we decide it, there is nothing else we can do about a . But what can we say about b ? We all know that hypothesis test cannot be foolproof; sometimes the test does not reject H0 when it happens to be a false one and this way a Type II error is made. But we would certainly like that b (the probability of accepting H0 when H0 is not true) to be as small as possible. Alternatively, we would like that 1 – b (the probability of rejecting H0 when H0 is not true) to be as large as possible. If 1 – b is very much nearer to unity (i.e., nearer to 1.0), we can infer that the test is working quite well, meaning thereby that the test is rejecting H0 when it is not true and if 1 – b is very much nearer to 0.0, then we infer that the test is poorly working, meaning thereby that it is not rejecting H0 when H0 is not true. Accordingly 1 – b value is the measure of how well the test is working or what is technically described as the power of the test. In case we plot the values of 1 – b for each possible value of the population parameter (say m , the true population mean) for which the H0 is not true (alternatively the Ha is true), the resulting curve is known as the power curve associated with the given test. Thus power curve of a hypothesis test is the curve that shows the conditional probability of rejecting H0 as a function of the population parameter and size of the sample.

The function defining this curve is known as the power function. In other words, the power function of a test is that function defined for all values of the parameter(s) which yields the probability that H0 is rejected and the value of the power function at a specific parameter point is called the power of the test at that point. As the population parameter gets closer and closer to hypothesized value of the population parameter, the power of the test (i.e., 1 – b) must get closer and closer to the probability of rejecting H0 when the population parameter is exactly equal to hypothesised value of the parameter. We know that this probability is simply the significance level of the test, and as such the power curve of a test terminates at a point that lies at a height of a (the significance level) directly over the population parameter.

Closely related to the power function, there is another function which is known as the operating characteristic function which shows the conditional probability of accepting H0 for all values of population parameter(s) for a given sample size, whether or not the decision happens to be a correct one. If power function is represented as H and operating characteristic function as L, then we have L = 1 – H.

8. Discuss the Importance of Parametric test in Research Methodology

The important parametric tests are:

a. z-test;
b.t-test;
c. χ^2-test, and
d. F-test.

All these tests are based on the assumption of normality i.e., the source of data is considered to be normally distributed.

In some cases the population may not be normally distributed, yet the tests will be applicable on account of the fact that we mostly deal with samples and the sampling distributions closely approach normal distributions.

z-test is based on the normal probability distribution and is used for judging the significance of several statistical measures, particularly the mean. The relevant test statistic, z, is worked out and compared with its probable value (to be read from table showing area under normal curve) at a specified level of significance for judging the significance of the measure concerned. This is a most frequently used test in research studies. This test is used even when binomial distribution or t-distribution is applicable on the presumption that such a distribution tends to approximate normal distribution as 'n' becomes larger. z-test is generally used for comparing the mean of a sample to some hypothesised mean for the population in case of large sample, or when population variance is known. z-test is also used for judging the significance of difference between means of two independent samples in case of large samples, or when population variance is known. z-test is also used for comparing the sample proportion to a theoretical value of population proportion or for judging the difference in proportions of two independent samples when n happens to be large. Besides, this test may be used for judging the significance of median, mode, coefficient of correlation and several other measures. t-test is based on t-distribution and is considered an appropriate test for judging the significance of a sample mean or for judging the significance of difference between the means of two samples in case of small sample(s) when population variance is not known (in which case we use variance of the sample as an estimate of the population variance). In case two samples are related, we use paired t-test (or what is known as difference test) for judging the significance of the mean of difference between the two related samples. It can also be used for judging the significance of the coefficients of simple and partial correlations. The relevant test statistic, t, is calculated from the sample data and then compared with its probable value based on t-distribution (to be read from the table that gives probable values of t for different levels of significance for different degrees of freedom) at a specified level of significance for concerning degrees of freedom for accepting or rejecting the null hypothesis. It may be noted that t-test applies only in case of small sample(s) when population variance is unknown.

χ^2-test is based on chi-square distribution and as a parametric test is used for comparing a sample variance to a theoretical population variance.

F-test is based on F-distribution and is used to compare the variance of the two-independent samples. This test is also used in the context of analysis of variance (ANOVA) for judging the significance of more than two sample means at one and the same time. It is also used for judging the significance of multiple correlation coefficients. Test statistic, F, is calculated and compared with its probable value (to be seen in the F-ratio tables for different degrees of freedom for greater and smaller variances at specified level of significance) for accepting or rejecting the null hypothesis.

9. Estimate the hypothesis of testing of mean in research methodology.

Mean of the population can be tested presuming different situations such as the population may be normal or other than normal, it may be finite or infinite, sample size may be large or small, variance of the population may be known or unknown and the alternative hypothesis may be two-sided or onesided. Our testing technique will differ in different situations. We may consider some of the important situations.

1. Population normal, population infinite, sample size may be large or small but variance of the population is known, Ha may be one-sided or two-sided:
   In such a situation z-test is used for testing hypothesis of mean and the test statistic z is worked our as under:

   $$z = \frac{\overline{X} - \mu_{H_0}}{\sigma_p/\sqrt{n}}$$

2. Population normal, population finite, sample size may be large or small but variance of the population is known, Ha may be one-sided or two-sided:
   In such a situation z-test is used and the test statistic z is worked out as under (using finite population multiplier):

   $$z = \frac{\overline{X} - \mu_{H_0}}{\left(\sigma_p/\sqrt{n}\right) \times \left[\sqrt{(N-n)/(N-1)}\right]}$$

3. Population normal, population infinite, sample size small and variance of the population unknown, Ha may be one-sided or two-sided:
   In such a situation t-test is used and the test statistic t is worked out as under:

   $$t = \frac{\overline{X} - \mu_{H_0}}{\sigma_s/\sqrt{n}} \text{ with d.f.} = (n-1)$$

   $$\sigma_s = \sqrt{\frac{\Sigma\left(X_i - \overline{X}\right)^2}{(n-1)}}$$

4. Population normal, population finite, sample size small and variance of the population unknown, and Ha may be one-sided or two-sided:
   In such a situation t-test is used and the test statistic ‘t’ is worked out as under (using finite population multiplier):

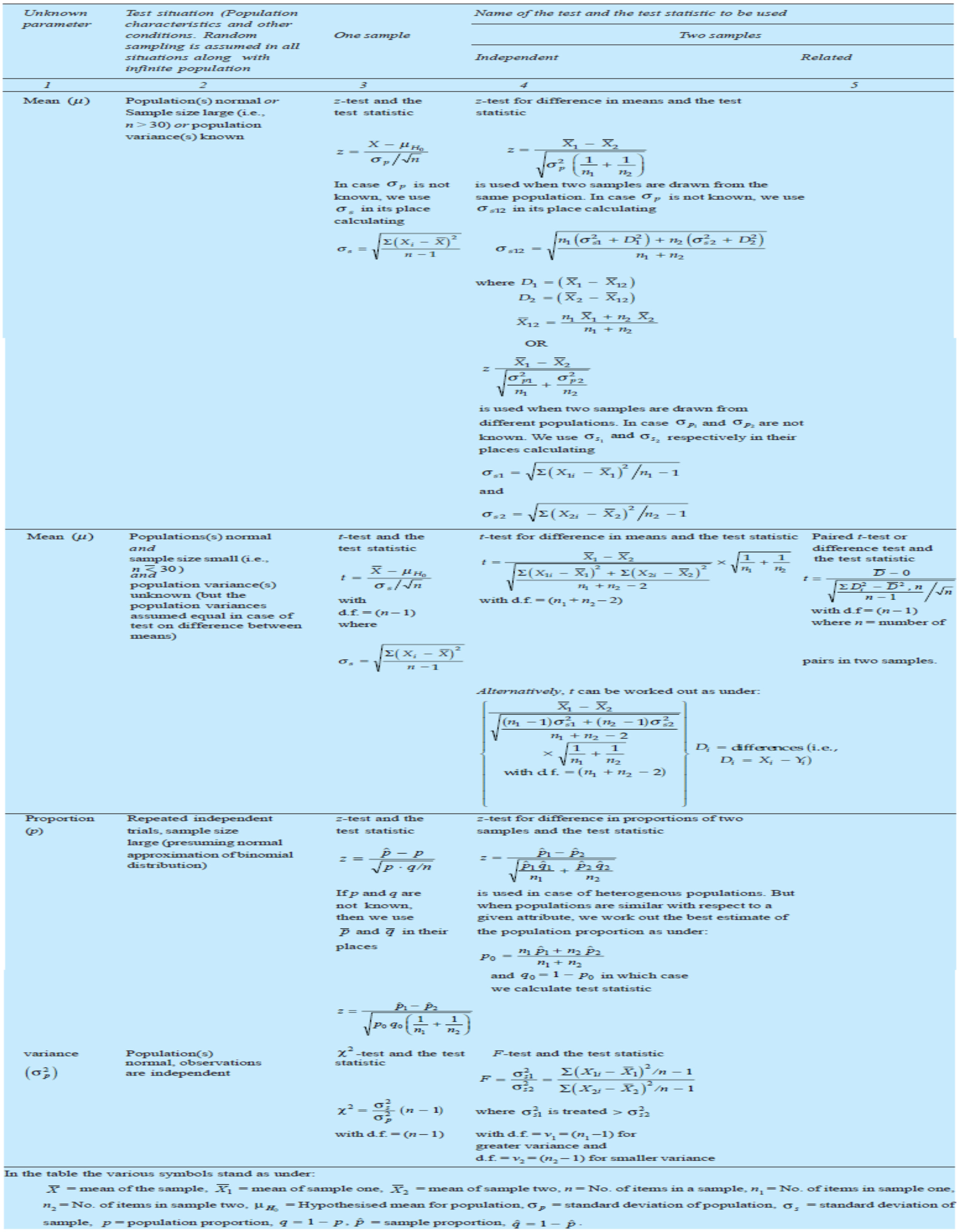
$$t = \frac{\overline{X} - \mu_{H_0}}{\left(\sigma_s/\sqrt{n}\right) \times \sqrt{(N-n)/(N-1)}} \text{ with d.f.} = (n-1)$$

5. Population may not be normal but sample size is large, variance of the population may be known or unknown, and Ha may be one-sided or two-sided:
   In such a situation we use z-test and work out the test statistic z as under:

   $$z = \frac{\overline{X} - \mu_{H_0}}{\sigma_p/\sqrt{n}}$$

   (This applies in case of infinite population when variance of the population is known but when variance is not known, we use ss in place of s p in this formula.)
   OR(This applies in case of finite population when variance of the population is known but when variance is not known, we use ss in place of s p in this formula.)

Table Names of Some Parametric Tests along with Test Situations and Test Statistics used in Context of Hypothesis Testing



Unknown parameter	*Test situation (Population characteristics and other conditions. Random sampling is assumed in all situations along with infinite population*	*One sample*	*Name of the test and the test statistic to be used*	
			Two samples	
			Independent	*Related*
1	*2*	*3*	*4*	*5*
Mean (μ)	Population(s) normal *or* Sample size large (i.e., $n > 30$) *or* population variance(s) known	z-test and the test statistic $z = \frac{X - \mu_{H_0}}{\sigma_p / \sqrt{n}}$ In case σ_p is not known, we use σ_s in its place calculating $\sigma_s = \sqrt{\frac{\Sigma(X_i - \overline{X})^2}{n-1}}$	z-test for difference in means and the test statistic $z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\sigma_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ is used when two samples are drawn from the same population. In case σ_p is not known, we use σ_{s12} in its place calculating $\sigma_{s12} = \sqrt{\frac{n_1(\sigma_{s1}^2 + D_1^2) + n_2(\sigma_{s2}^2 + D_2^2)}{n_1 + n_2}}$ where $D_1 = (\overline{X}_1 - \overline{X}_{12})$ $D_2 = (\overline{X}_2 - \overline{X}_{12})$ $\overline{X}_{12} = \frac{n_1 \overline{X}_1 + n_2 \overline{X}_2}{n_1 + n_2}$ OR $z \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_{p1}^2}{n_1} + \frac{\sigma_{p2}^2}{n_2}}}$ is used when two samples are drawn from different populations. In case σ_{p_1} and σ_{p_2} are not known. We use σ_{s_1} and σ_{s_2} respectively in their places calculating $\sigma_{s1} = \sqrt{\Sigma(X_{1i} - \overline{X}_1)^2 / n_1 - 1}$ and $\sigma_{s2} = \sqrt{\Sigma(X_{2i} - \overline{X}_2)^2 / n_2 - 1}$	
Mean (μ)	Populations(s) normal *and* sample size small (i.e., $n \lesssim 30$) *and* population variance(s) unknown (but the population variances assumed equal in case of test on difference between means)	t-test and the test statistic $t = \frac{\overline{X} - \mu_{H_0}}{\sigma_s / \sqrt{n}}$ with d.f. $= (n-1)$ where $\sigma_s = \sqrt{\frac{\Sigma(X_i - \overline{X})^2}{n-1}}$	t-test for difference in means and the test statistic $t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\Sigma(X_{1i} - \overline{X}_1)^2 + \Sigma(X_{2i} - \overline{X}_2)^2}{n_1 + n_2 - 2}}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ with d.f. $= (n_1 + n_2 - 2)$ *Alternatively*, t can be worked out as under: $\left\{ \begin{array}{c} \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{(n_1 - 1)\sigma_{s1}^2 + (n_2 - 1)\sigma_{s2}^2}{n_1 + n_2 - 2}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ \text{with d.f.} = (n_1 + n_2 - 2) \end{array} \right\}$	Paired t-test or difference test and the test statistic $t = \frac{\overline{D} - 0}{\sqrt{\frac{\Sigma D_i^2 - \overline{D}^2 . n}{n-1}} \Big/ \sqrt{n}}$ with d.f $= (n-1)$ where $n =$ number of pairs in two samples. $D_i =$ differences (i.e., $D_i = X_i - Y_i$)
Proportion (p)	Repeated independent trials, sample size large (presuming normal approximation of binomial distribution)	z-test and the test statistic $z = \frac{\hat{p} - p}{\sqrt{p \cdot q / n}}$ If p and q are not known, then we use $\overline{p}$ and $\overline{q}$ in their places $z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p_0 q_0 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$	z-test for difference in proportions of two samples and the test statistic $z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}}$ is used in case of heterogenous populations. But when populations are similar with respect to a given attribute, we work out the best estimate of the population proportion as under: $p_0 = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$ and $q_0 = 1 - p_0$ in which case we calculate test statistic	
variance (σ_p^2)	Population(s) normal, observations are independent	χ^2-test and the test statistic $\chi^2 = \frac{\sigma_s^2}{\sigma_p^2}(n-1)$ with d.f. $= (n-1)$	F-test and the test statistic $F = \frac{\sigma_{s1}^2}{\sigma_{s2}^2} = \frac{\Sigma(X_{1i} - \overline{X}_1)^2 / n - 1}{\Sigma(X_{2i} - \overline{X}_2)^2 / n - 1}$ where σ_{s1}^2 is treated $> \sigma_{s2}^2$ with d.f. $= v_1 = (n_1 - 1)$ for greater variance and d.f. $= v_2 = (n_2 - 1)$ for smaller variance	

In the table the various symbols stand as under:

$\overline{X}$ = mean of the sample, $\overline{X}_1$ = mean of sample one, $\overline{X}_2$ = mean of sample two, n = No. of items in a sample, n_1 = No. of items in sample one, n_2 = No. of items in sample two, μ_{H_0} = Hypothesised mean for population, σ_p = standard deviation of population, σ_s = standard deviation of sample, p = population proportion, $q = 1 - p$, $\hat{p}$ = sample proportion, $\hat{q} = 1 - \hat{p}$.

10. Discuss testing for differences between means in research methodology.

In many decision-situations, we may be interested in knowing whether the parameters of two populations are alike or different. For instance, we may be interested in testing whether female workers earn less than male workers for the same job. We shall explain now the technique of hypothesis testing for differences between means. The null hypothesis for testing of difference between means is generally stated as H0 : m1 = m2 , where m1 is population mean of one population and m2 is population mean of the second population, assuming both the populations to be normal populations. Alternative hypothesis may be of not equal to or less than or greater than type as stated earlier and accordingly we shall determine the acceptance or rejection regions for testing the hypotheses. There may be different situations when we are examining the significance of difference between two means, but the following may be taken as the usual situations:

1. Population variances are known or the samples happen to be large samples:
   In this situation we use z-test for difference in means and work out the test statistic z as under:

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_{p1}^2}{n_1} + \frac{\sigma_{p2}^2}{n_2}}}$$

   In case σ_{p_1} and σ_{p_2} are not known, we use σ_{s_1} and σ_{s_2} respectively in their places calculating

$$\sigma_{s_1} = \sqrt{\frac{\Sigma\left(X_{1i} - \bar{X}_1\right)^2}{n_1 - 1}} \text{ and } \sigma_{s_2} = \sqrt{\frac{\Sigma\left(X_{2i} - \bar{X}_2\right)^2}{n_2 - 1}}$$

2. Samples happen to be large but presumed to have been drawn from the same population whose variance is known:
   In this situation we use z test for difference in means and work out the test statistic z as Under

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

   In case σ_p is not known, we use $\sigma_{s_{1.2}}$ (combined standard deviation of the two samples) in its place calculating

$$\sigma_{s_{1.2}} = \sqrt{\frac{n_1\left(\sigma_{s_1}^2 + D_1^2\right) + n_2\left(\sigma_{s_2}^2 + D_2^2\right)}{n_1 + n_2}}$$

   where $D_1 = \left(\bar{X}_1 - \bar{X}_{1.2}\right)$

   $D_2 = \left(\bar{X}_2 - \bar{X}_{1.2}\right)$

3. Samples happen to be small samples and population variances not known but assumed to be equal:
   In this situation we use t-test for difference in means and work out the test statistic t as

under:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\Sigma\left(X_{1i} - \bar{X}_1\right)^2 + \Sigma\left(X_{2i} - \bar{X}_2\right)^2}{n_1 + n_2 - 2}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

with d.f. $= (n_1 + n_2 - 2)$

Alternatively, we can also state

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1 - 1)\sigma_{s_1}^2 + (n_2 - 1)\sigma_{s_2}^2}{n_1 + n_2 - 2}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

with d.f. $= (n_1 + n_2 - 2)$

11. Write a sort note on hypothesis testing for comparing two related samples.

Paired t-test is a way to test for comparing two related samples, involving small values of n that does not require the variances of the two populations to be equal, but the assumption that the two populations are normal must continue to apply. For a paired t-test, it is necessary that the observations in the two samples be collected in the form of what is called matched pairs i.e., “each observation in the one sample must be paired with an observation in the other sample in such a manner that these observations are somehow “matched” or related, in an attempt to eliminate extraneous factors which are not of interest in test.”5Such a test is generally considered appropriate in a before-and-after-treatment study. For instance, we may test a group of certain students before and after training in order to know whether the training is effective, in which situation we may use paired t-test. To apply this test, we first work out the difference score for each matched pair, and then find out the average of such differences, D , along with the sample variance of the difference score. If the values from the two matched samples are denoted as Xi and Yi and the differences by Di (Di = Xi – Yi), then the mean of the differences i.e.,

$$\bar{D} = \frac{\sum D_i}{n}$$

and the variance of the differences or

$$\left(\sigma_{diff.}\right)^2 = \frac{\sum D_i^2 - \left(\bar{D}\right)^2 \cdot n}{n - 1}$$

Assuming the said differences to be normally distributed and independent, we can apply the paired t-test for judging the significance of mean of differences and work out the test statistic t as under:

$t = \frac{\bar{D} - 0}{\sigma_{diff} / \sqrt{n}}$ with $(n - 1)$ degrees of freedom

12. Discuss the limitations of hypothesis testing in research.

We have described above some important test often used for testing hypotheses on the basis of which important decisions may be based. But there are several limitations of the said tests which should always be borne in mind by a researcher. Important limitations are as follows:

1. The tests should not be used in a mechanical fashion. It should be kept in view that testing is not decision-making itself; the tests are only useful aids for decision-making. Hence “proper interpretation of statistical evidence is important to intelligent decisions.”

2. Test do not explain the reasons as to why does the difference exist, say between the means of the two samples. They simply indicate whether the difference is due to fluctuations of sampling or because of other reasons but the tests do not tell us as to which is/are the other reason(s) causing the difference.

3. Results of significance tests are based on probabilities and as such cannot be expressed with full certainty. When a test shows that a difference is statistically significant, then it simply suggests that the difference is probably not due to chance.

4. Statistical inferences based on the significance tests cannot be said to be entirely correct evidences concerning the truth of the hypothesis. This is specially so in case of small samples where the probability of drawing erring inferences happens to be generally higher. For greater reliability, the size of samples be sufficiently enlarged.

All these limitations suggest that in problems of statistical significance, the inference techniques (or the tests) must be combined with adequate knowledge of the subject-matter along with the ability of good judgement.

12. Note on Chi-square test for comparing variable

The chi-square test of variance is often used to judge the significance of population variance i.e., we can use the test to judge if a random sample has been drawn from a normal population with mean (m) and with a specified variance (s p 2). The test is based on c2 -distribution. Such a distribution we encounter when we deal with collections of values that involve adding up squares. Variances of samples require us to add a collection of squared quantities and, thus, have distributions that are related to c2 -distribution. If we take each one of a collection of sample variances, divided them by the known population variance and multiply these quotients by (n – 1), where n means the number of items in the sample, we shall obtain a c2 -distribution. Thus, Formula degrees of freedom.

The x2 -distribution is not symmetrical and all the values are positive. For making use of this distribution, one is required to know the degrees of freedom since for different degrees of freedom we have different curves. The smaller the number of degrees of freedom, the more skewed is the distribution which is illustrated in below:

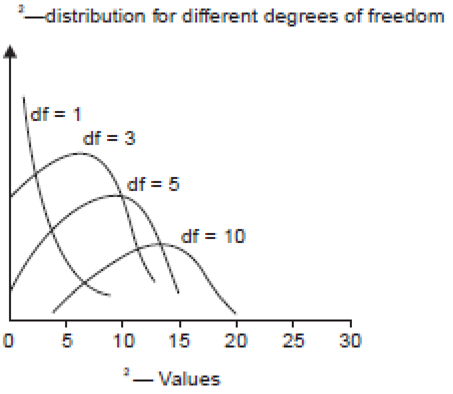


Table given in the Appendix gives selected critical values of x2 for the different degrees of freedom. x2 -values are the quantities indicated on the x-axis of the above diagram and in the table are areas below that value.
In brief, when we have to use chi-square as a test of population variance, we have to work out the value of c2 to test the null hypothesis (viz., Formula) as under:

$$\chi^2 = \frac{\sigma_s^2}{\sigma_p^2}(n-1)$$

Then by comparing the calculated value with the table value of c2 for (n – 1) degrees of freedom at a given level of significance, we may either accept or reject the null hypothesis. If the calculated value of c2 is less than the table value, the null hypothesis is accepted, but if the calculated value is equal or greater than the table value, the hypothesis is rejected.

13. When we apply Chi-square test in social science research?

The following conditions should be satisfied before x2 test can be applied:

1. Observations recorded and used are collected on a random basis.

2. All the itmes in the sample must be independent.

3. No group should contain very few items, say less than 10. In case where the frequencies are less than 10, regrouping is done by combining the frequencies of adjoining groups so that the new frequencies become greater than 10. Some statisticians take this number as 5, but 10 is regarded as better by most of the statisticians.

4. The overall number of items must also be reasonably large. It should normally be at least 50, howsoever small the number of groups may be.

5. The constraints must be linear. Constraints which involve linear equations in the cell frequencies of a contingency table (i.e., equations containing no squares or higher powers of the frequencies) are known are know as linear constraints.

14. What are the steps involved in applying Chi-square test in social science research.

The various chi square test steps involved are as follows:

1. First of all calculate the expected frequencies on the basis of given hypothesis or on the basis of null hypothesis. Usually in case of a 2 × 2 or any contingency table, the expected frequency for any given cell is worked out as under:

$$\text{Expected frequency of any cell} = \left[\frac{\begin{array}{c}\text{(Row total for the row of that cell)} \times \\ \text{(Column total for the column of that cell)}\end{array}}{\text{(Grand total)}}\right]$$

2. Obtain the difference between observed and expected frequencies and find out the squares of such differences i.e., calculate (Oij – Eij)2.

3. Divide the quantity (Oij – Eij)2 obtained as stated above by the corresponding expected frequency to get (Oij – Eij)2/Eij and this should be done for all the cell frequencies or the group frequencies.

4. Find the summation of (Oij – Eij)2/Eij values or what we call $\chi^2 = \Sigma \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$. This is the required x2 value.

The 2x2 value obtained as such should be compared with relevant table value of x2 and then inference be drawn as stated above.

15. Note on important characteristics of Chi-square test in Statistics.

The Characterstics of Chi square test in statiscs are given below

1. This test (as a non-parametric test) is based on frequencies and not on the parameters like mean and standard deviation.
2. The test is used for testing the hypothesis and is not useful for estimation.
3. This test possesses the additive property as has already been explained.
4. This test can also be applied to a complex contingency table with several classes and as such is a very useful test in research work.
5. This test is an important non-parametric test as no rigid assumptions are necessary in regard to the type of population, no need of parameter values and relatively less mathematical details are involved.

16. **Note on caution in using Chi-square test**

The chi-square test is no doubt a most frequently used test, but its correct application is equally an uphill task. It should be borne in mind that the test is to be applied only when the individual observations of sample are independent which means that the occurrence of one individual observation (event) has no effect upon the occurrence of any other observation (event) in the sample under consideration. Small theoretical frequencies, if these occur in certain groups, should be dealt with under special care. The other possible reasons concerning the improper application or misuse of this test can be

a. neglect of frequencies of non-occurrence;
b. failure to equalise the sum of observed and the sum of the expected frequencies;
c. wrong determination of the degrees of freedom;
d. wrong computations, and the like.

The researcher while applying this test must remain careful about all these things and must thoroughly understand the rationale of this important test before using it and drawing inferences in respect of his hypothesis.

ANOVA

17. What is ANOVA?

Analysis of variance (abbreviated as ANOVA) is an extremely useful technique concerning researches in the fields of economics, biology, education, psychology, sociology, business/industry and in researches of several other disciplines. This technique is used when multiple sample cases are involved. As stated earlier, the significance of the difference between the means of two samples can be judged through either z-test or the t-test, but the difficulty arises when we happen to examine the significance of the difference amongst more than two sample means at the same time. The ANOVA technique enables us to perform this simultaneous test and as such is considered to be an important tool of analysis in the hands of a researcher. Using this technique, one can draw inferences about whether the samples have been drawn from populations having the same mean.

18. Why ANOVA test is important in Research Methodology

The ANOVA is important in the context of all those situations where we want to compare more than two populations such as in comparing the yield of crop from several varieties of seeds, the gasoline mileage of four automobiles, the smoking habits of five groups of university students and so on. In such circumstances one generally does not want to consider all possible combinations of two populations at a time for that would require a great number of tests before we would be able to arrive at a decision. This would also consume lot of time and money, and even then certain relationships may be left unidentified (particularly the interaction effects). Therefore, one quite often utilizes the ANOVA technique and through it investigates the differences among the means of all the populations simultaneously.

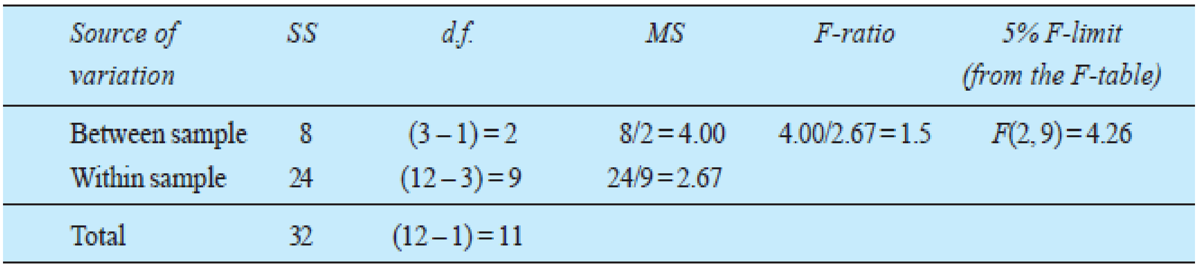
19. Describe the basic principles of ANOVA.

The basic principle of ANOVA is to test for differences among the means of the populations by examining the amount of variation within each of these samples, relative to the amount of variation between the samples. In terms of variation within the given population, it is assumed that the values of (Xij) differ from the mean of this population only because of random effects i.e., there are influences on (Xij) which are unexplainable, whereas in examining differences between populations we assume that the difference between the mean of the jth population and the grand mean is attributable to what is called a 'specific factor' or what is technically described as treatment effect. Thus while using ANOVA, we assume that each of the samples is drawn from a normal population and that each of these populations has the same variance. We also assume that all factors other than the one or more being tested are effectively controlled. This, in other words, means that we assume the absence of many factors that might affect our conclusions concerning the factor(s) to be studied.
In short, we have to make two estimates of population variance viz., one based on between samples variance and the other based on within samples variance. Then the said two estimates of population variance are compared with F-test, wherein we work out.
F=**Estimate of population variance based on between samples variance / Estimate of population variance based on within samples variance**

This value of F is to be compared to the F-limit for given degrees of freedom. If the F value we work out is equal or exceeds* the F-limit value, we may say that there are significant differences between the sample means.

20. What is Coding Method?

Coding method is furtherance of the short-cut method. This is based on an important property of F-ratio that its value does not change if all the n item values are either multiplied or divided by a common figure or if a common figure is either added or subtracted from each of the given n item values. Through this method big figures are reduced in magnitude by division or subtraction and computation work is simplified without any disturbance on the F-ratio. This method should be used specially when given figures are big or otherwise inconvenient. Once the given figures are converted with the help of some common figure, then all the steps of the short-cut method stated above can be adopted for obtaining and interpreting F-ratio.



Source of variation	*SS*	*d.f.*	*MS*	*F-ratio*	*5% F-limit (from the F-table)*
Between sample	8	$(3-1)=2$	$8/2=4.00$	$4.00/2.67=1.5$	$F(2, 9)=4.26$
Within sample	24	$(12-3)=9$	$24/9=2.67$		
Total	32	$(12-1)=11$			

Multivariate Analysis Techniques

21. What is multivariate analysis technique?

Multivariate techniques have emerged as a powerful tool to analyse data represented in terms of many variables. The main reason being that a series of univariate analysis carried out separately for each variable may, at times, lead to incorrect interpretation of the result. This is so because univariate analysis does not consider the correlation or inter-dependence among the variables. As a result, during the last fifty years, a number of statisticians have contributed to the development of several multivariate techniques. Today, these techniques are being applied in many fields such as economics, sociology, psychology, agriculture, anthropology, biology and medicine. These techniques are used in analyzing social, psychological, medical and economic data, specially when the variables concerning research studies of these fields are supposed to be correlated with each other and when rigorous probabilistic models cannot be appropriately used. Applications of multivariate techniques in practice have been accelerated in modern times because of the advent of high speed electronic computers.

22. Evaluate the characteristics of Multivariate analysis techniques

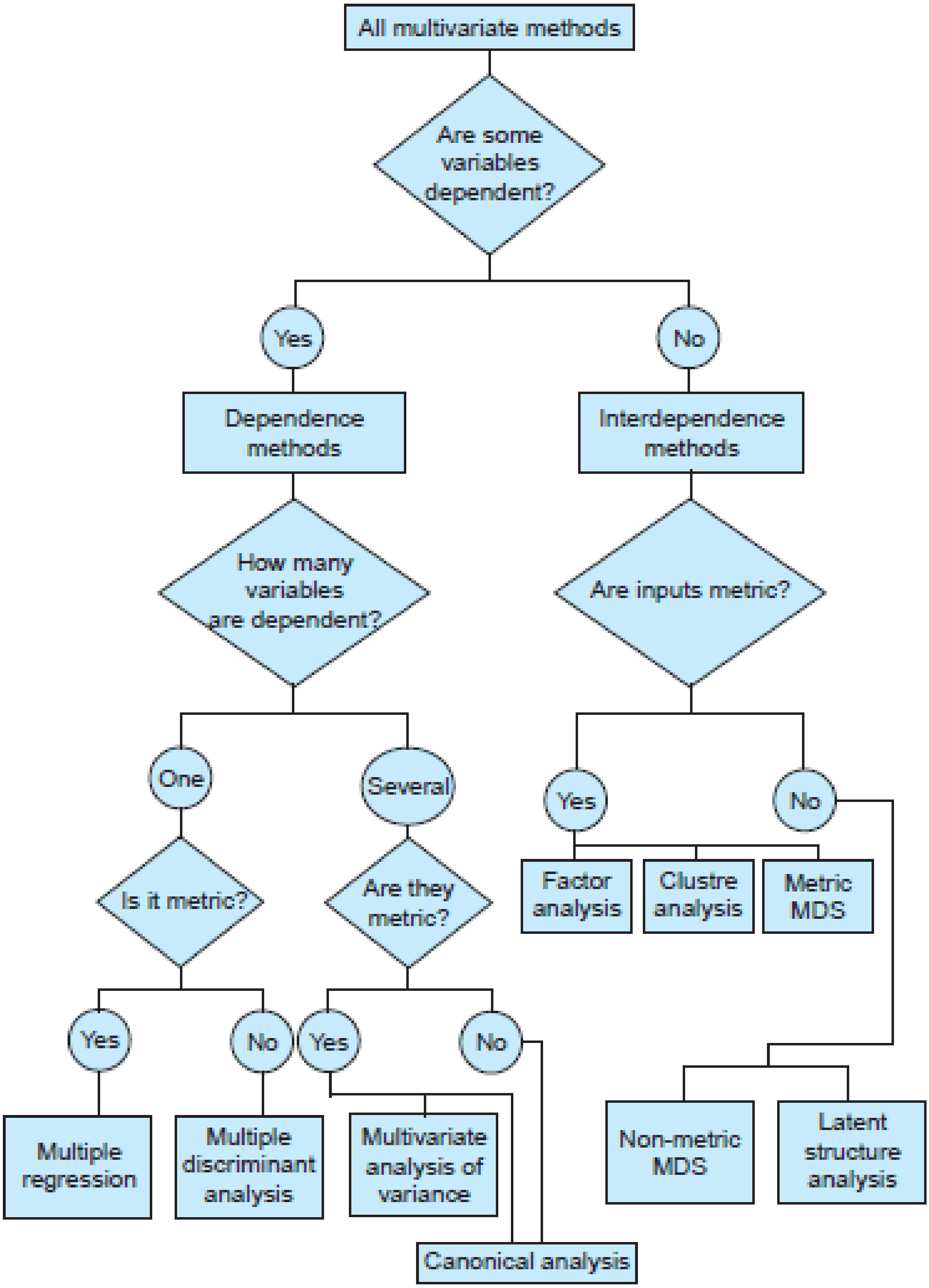
Multivariate analysis techniques are largely empirical and deal with the reality; they possess the ability to analyse complex data. Accordingly in most of the applied and behavioural researches, we generally resort to multivariate analysis techniques for realistic results. Besides being a tool for analyzing the data, multivariate techniques also help in various types of decision-making. For example, take the case of college entrance examination wherein a number of tests are administered to candidates, and the candidates scoring high total marks based on many subjects are admitted. This system, though apparently fair, may at times be biased in favour of some subjects with the larger standard deviations. Multivariate techniques may be appropriately used in such situations for developing norms as to who should be admitted in college. We may also cite an example from medical field. Many medical examinations such as blood pressure and cholesterol tests are administered to patients. Each of the results of such examinations has significance of its own, but it is also important to consider relationships between different test results or results of the same tests at different occasions in order to draw proper diagnostic conclusions and to determine an appropriate therapy. Multivariate techniques can assist us in such a situation. In view of all this, we can state that "if the researcher is interested in making probability statements on the basis of sampled multiple measurements, then the best strategy of data analysis is to use some suitable multivariate statistical technique.”

The basic objective underlying multivariate techniques is to represent a collection of massive data in a simplified way. In other words, multivariate techniques transform a mass of observations into a smaller number of composite scores in such a way that they may reflect as much information as possible contained in the raw data obtained concerning a research study. Thus, the main contribution of these techniques is in arranging a large amount of complex information involved in the real data into a simplified visible form. Mathematically, multivariate techniques consist in “forming a linear composite vector in a vector subspace, which can be represented in terms of projection of a vector onto certain specified subspaces.”

For better appreciation and understanding of multivariate techniques, one must be familiar with fundamental concepts of linear algebra, vector spaces, orthogonal and oblique projections and univariate analysis. Even then before applying multivariate techniques for meaningful results, one must consider the nature and structure of the data and the real aim of the analysis. We should also not forget that multivariate techniques do involve several complex mathematical computations and as such can be utilized largely with the availability of computer facility.

23. Describe the classification of Multivariate Analysis Techniques in Research Methodology

Multivariate analysis techniques which can be conveniently classified into two broad categories viz., dependence methods and interdependence methods. This sort of classification depends upon the question: Are some of the involved variables dependent upon others? If the answer is ‘yes’, we have dependence methods; but in case the answer is ‘no’, we have interdependence methods. Two more questions are relevant for understanding the nature of multivariate techniques. Firstly, in case some variables are dependent, the question is how many variables are dependent? The other question is, whether the data are metric or non-metric? This means whether the data are quantitative, collected on interval or ratio scale, or whether the data are qualitative, collected on nominal or ordinal scale. The technique to be used for a given situation depends upon the answers to all these very questions. Jadish N. Sheth in his article on “The multivariate revolution in marketing research” has given the flow chart that clearly exhibits the nature of some important multivariate techniques as shown in Fig. below. Thus, we have two types of multivariate techniques: one type for data containing both dependent and independent variables, and the other type for data containing several variables without dependency relationship. In the former category are included techniques like multiple regression analysis, multiple discriminant analysis, multivariate analysis of variance and canonical analysis, whereas in the latter category we put techniques like factor analysis, cluster analysis, multidimensional scaling or MDS (both metric and non-metric) and the latent structure analysis.


All multivariate methods
Are some variables dependent?
Yes
No
Dependence methods
Interdependence methods
How many variables are dependent?
Are inputs metric?
One
Several
Yes
No
Is it metric?
Are they metric?
Factor analysis
Clustre analysis
Metric MDS
Yes
No
Yes
No
Multiple regression
Multiple discriminant analysis
Multivariate analysis of variance
Non-metric MDS
Latent structure analysis
Canonical analysis

24. Describe the variables of Multivariate analysis.

The various multivariate techniques, it seems appropriate to have a clear idea about the term, ‘variables’ used in the context of multivariate analysis. Many variables used in multivariate analysis can be classified into different categories from several points of view. Important ones are as under:

1. Explanatory variable and criterion variable: If X may be considered to be the cause of Y, then X is described as explanatory variable (also termed as causal or independent variable) and Y is described as criterion variable (also termed as resultant or dependent variable). In some cases both explanatory variable and criterion variable may consist of a set of many variables in which case set (X1, X2, X3, …., Xp) may be called a set of explanatory variables and the set (Y1, Y2, Y3, …., Yq) may be called a set of criterion variables if the variation of the former may be supposed to cause the variation of the latter as a whole. In economics, the explanatory variables are called external or exogenous variables and the criterion variables are called endogenous variables. Some people use the term external criterion for explanatory variable and the term internal criterion for criterion variable.
2. Observable variables and latent variables: Explanatory variables described above are supposed to be observable directly in some situations, and if this is so, the same are termed as observable variables. However, there are some unobservable variables which may influence the criterion variables. We call such unobservable variables as latent variables.
3. Discrete variable and continuous variable: Discrete variable is that variable which when measured may take only the integer value whereas continuous variable is one which, when measured, can assume any real value (even in decimal points).
4. Dummy variable (or Pseudo variable): This term is being used in a technical sense and is useful in algebraic manipulations in context of multivariate analysis. We call Xi (i = 1, …., m) a dummy variable, if only one of Xi is 1 and the others are all zero.

25. Describe the various multivariate analysis.

A brief description of the various multivariate techniques named above (with special emphasis on factor analysis) is as under:

1. ***Multiple regression:*** In multiple regression we form a linear composite of explanatory variables in such way that it has maximum correlation with a criterion variable. This technique is appropriate when the researcher has a single, metric criterion variable. Which is supposed to be a function of other explanatory variables. The main objective in using this technique is to predict the variability the dependent variable based on its covariance with all the independent variables. One can predict the level of the dependent phenomenon through multiple regression analysis model, given the levels of independent variables. Given a dependent variable, the linear-multiple regression problem is to estimate constants B1, B2, ... Bk and A such that the expression Y = B1X1 + B2X2 + ... + BkXk + A pare rovides a good estimate of an individual’s Y score based on his X scores.
   In practice, Y and the several X variables are converted to standard scores; zy, zl, z2, ... zk; each z has a mean of 0 and standard deviation of 1. Then the problem is to estimate constants, bi , such that

z¢y = b1z1 + b2z2 + ...+ bk zk

where z'y stands for the predicted value of the standardized Y score, zy. The expression on the right side of the above equation is the linear combination of explanatory variables. The constant A is eliminated in the process of converting X’s to z’s. The least-squares-method is used, to estimate the beta weights in such a way that the sum of the squared prediction errors is kept as small as possible i.e., the Formula is minimized. The predictive adequacy of a set of beta weights is indicated by the size of the correlation coefficient rzy × z¢y between the predicted z¢y scores and the actual zy scores. This special correlation coefficient from Karl Pearson is termed the multiple correlation coefficient (R). The squared multiple correlation, R2, represents the proportion of criterion (zy) variance accounted for by the explanatory variables, i.e., the proportion of total variance that is ‘Common Variance’.

Sometimes the researcher may use step-wise regression techniques to have a better idea of the independent contribution of each explanatory variable. Under these techniques, the investigator adds the independent contribution of each explanatory variable into the prediction equation one by one, computing betas and R2 at each step. Formal computerized techniques are available for the purpose and the same can be used in the context of a particular problem being studied by the researcher.

2. **Multiple discriminant analysis:** Through discriminant analysis technique, researcher may classify individuals or objects into one of two or more mutually exclusive and exhaustive groups on the basis of a set of independent variables. Discriminant analysis requires interval independent variables and a nominal dependent variable. For example, suppose that brand preference (say brand x or y) is the dependent variable of interest and its relationship to an individual’s income, age, education, etc. is being investigated, then we should use the technique of discriminant analysis. Regression analysis in such a situation is not suitable because the dependent variable is, not intervally scaled. Thus discriminant analysis is considered an appropriate technique when the single dependent variable happens to be non-metric and is to be classified into two or more groups, depending upon its relationship with several independent variables which all happen to be metric. The objective in discriminant analysis happens to be to predict an object’s likelihood of belonging to a particular group based on several independent variables. In case we classify the dependent variable in more than two groups, then we use the name multiple discriminant analysis; but in case only two groups are to be formed, we simply use the term discriminant analysis. We may briefly refer to the technical aspects relating to discriminant analysis.
3. There happens to be a simple scoring system that assigns a score to each individual or object. This score is a weighted average of the individual’s numerical values of his independent variables. On the basis of this score, the individual is assigned to the ‘most likely’ category. For example, an individual is 20 years old, has an annual income of Rs 12,000,and has 10 years of formal education. Let b1, b2, and b3 be the weights attached to the independent variables of age, income and education respectively. The individual’s score (z), assuming linear score, would be:

   z = b1 (20) + b2 (12000) + b3 (10)

   This numerical value of z can then be transformed into the probability that the individual is an early user, a late user or a non-user of the newly marketed consumer product (here we are making three categories viz. early user, late user or a non-user).
4. The numerical values and signs of the b’s indicate the importance of the independent variables in their ability to discriminate among the different classes of individuals. Thus, through the discriminant analysis, the researcher can as well determine which independent variables are most useful in predicting whether the respondent is to be put into one group or the other. In other words, discriminant analysis reveals which specific variables in the profile account for the largest proportion of inter-group differences.
5. In case only two groups of the individuals are to be formed on the basis of several independent variables, we can then have a model like this

   zi = b0 + b1X1i + b2X2i + ... + bnXni

where Xji = the ith individual's value of the jth independent variable;
bj = the discriminant coefficient of the jth variable;
zi = the ith individual's discriminant score;
zcrit. = the critical value for the discriminant score.
The classification procedure in such a case would be
If zi > zcrit., classify individual i as belonging to Group I
If zi < zcrit, classify individual i as belonging to Group II.
When n (the number of independent variables) is equal to 2, we have a straight line classification boundary. Every individual on one side of the line is classified as Group I and on the other side, every one is classified as belonging to Group II. When n = 3, the classification boundary is a two-dimensional plane in 3 space and in general the classification boundary is an n – 1 dimensional hyper-plane in n space.

6. In n-group discriminant analysis, a discriminant function is formed for each pair of groups. If there are 6 groups to be formed, we would have 6(6 – 1)/2 = 15 pairs of groups, and hence 15 discriminant functions. The b values for each function tell which variables are important for discriminating between particular pairs of groups. The z score for each discriminant function tells in which of these two groups the individual is more likely to belong. Then use is made of the transitivity of the relation "more likely than". For example, if group II is more likely than group I and group III is more likely than group II, then group III is also more likely than group I. This way all necessary comparisons are made and the individual is assigned to the most likely of all the groups. Thus, the multiple-group discriminant analysis is just like the two-group discriminant analysis for the multiple groups are simply examined two at a time.
7. For judging the statistical significance between two groups, we work out the Mahalanobis statistic, D2, which happens to be a generalized distance between two groups, where each group is characterized by the same set of n variables and where it is assumed that variancecovariance structure is identical for both groups. It is worked out thus:

$$D^2 = (U_1 - U_2)v^{-1}(U_1 - U_2)'$$

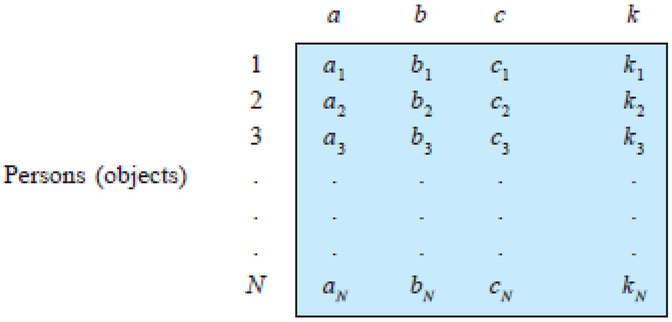
where U1 = the mean vector for group I
U2 = the mean vector for group II
v = the common variance matrix
By transformation procedure, this D2 statistic becomes an F statistic which can be used to see if the two groups are statistically different from each other.
From all this, we can conclude that the discriminant analysis provides a predictive equation, measures the relative importance of each variable and is also a measure of the ability of the equation to predict actual class-groups (two or more) concerning the dependent variable.

8. **Multivariate analysis of variance:** Multivariate analysis of variance is an extension of bivariate analysis of variance in which the ratio of among-groups variance to within-groups variance is calculated on a set of variables instead of a single variable. This technique is considered appropriate when several metric dependent variables are involved in a research study along with many non-metric explanatory variables. (But if the study has only one metric dependent variable and several nonmetric explanatory variables, then we use the ANOVA technique as explained earlier in the book.) In other words, multivariate analysis of variance is specially applied whenever the researcher wants to test hypotheses concerning multivariate differences in group responses to experimental manipulations. For instance, the market researcher may be interested in using one test market and one control market to examine the effect of an advertising campaign on sales as well as awareness, knowledge and attitudes. In that case he should use the technique of multivariate analysis of variance for meeting his objective.
9. **Canonical correlation analysis:** This technique was first developed by Hotelling wherein an effort is made to simultaneously predict a set of criterion variables from their joint co-

variance with a set of explanatory variables. Both metric and non-metric data can be used in the context of this multivariate technique. The procedure followed is to obtain a set of weights for the dependent and independent variables in such a way that linear composite of the criterion variables has a maximum correlation with the linear composite of the explanatory variables. For example, if we want to relate grade school adjustment to health and physical maturity of the child, we can then use canonical correlation analysis, provided we have for each child a number of adjustment scores (such as tests, teacher's ratings, parent's ratings and so on) and also we have for each child a number of health and physical maturity scores (such as heart rate, height, weight, index of intensity of illness and so on). The main objective of canonical correlation analysis is to discover factors separately in the two sets of variables such that the multiple correlation between sets of factors will be the maximum possible. Mathematically, in canonical correlation analysis, the weights of the two sets viz., a1, a2, … ak and yl, y2, y3, ... yj are so determined that the variables X = a1X1 + a2X2 +... + akXk + a and Y = y1Y1 + y2Y2 + … yjYj + y have a maximum common variance. The process of finding the weights requires factor analyses with two matrices.* The resulting canonical correlation solution then gives an over all description of the presence or absence of a relationship between the two sets of variables.

10. **Factor analysis:** Factor analysis is by far the most often used multivariate technique of research studies, specially pertaining to social and behavioural sciences. It is a technique applicable when there is a systematic interdependence among a set of observed or manifest variables and the researcher is interested in finding out something more fundamental or latent which creates this commonality.For instance, we might have data, say, about an individual's income, education, occupation and dwelling area and want to infer from these some factor (such as social class) which summarises the commonality of all the said four variables. The technique used for such purpose is generally described as factor analysis. Factor analysis, thus, seeks to resolve a large set of measured variables in terms of relatively few categories, known as factors. This technique allows the researcher to group variables into factors (based on correlation between variables) and the factors so derived may be treated as new variables (often termed as latent variables) and their value derived by summing the values of the original variables which have been grouped into the factor. The meaning and name of such new variable is subjectively determined by the researcher. Since the factors happen to be linear combinations of data, the coordinates of each observation or variable is measured to obtain what are called factor loadings. Such factor loadings represent the correlation between the particular variable and the factor, and are usually place in a matrix of correlations between the variable and the factors.
The mathematical basis of factor analysis concerns a data matrix* (also termed as score matrix), symbolized as S. The matrix contains the scores of N persons of k measures. Thus a1 is the score of person 1 on measure a, a2 is the score of person 2 on measure a, and kN is the score of person N on measure k. The score matrix then take the form as shown following:

SCORE MATRIX (or Matrix S)



$$\begin{array}{ll|cccc|} & & a & b & c & k \\ & 1 & a_1 & b_1 & c_1 & k_1 \\ & 2 & a_2 & b_2 & c_2 & k_2 \\ & 3 & a_3 & b_3 & c_3 & k_3 \\ \text{Persons (objects)} & . & . & . & . & . \\ & . & . & . & . & . \\ & . & . & . & . & . \\ & N & a_N & b_N & c_N & k_N \end{array}$$

It is assumed that scores on each measure are standardized [i.e., xi = (X - Xi)2 /si] . This being so, the sum of scores in any column of the matrix, S, is zero and the variance of scores in any column is 1.0. Then factors (a factor is any linear combination of the variables in a data matrix and can be stated in a general way like: A = Waa + Wbb + … + Wkk) are obtained (by any method of factoring). After this, we work out factor loadings (i.e., factor-variable correlations). Then communality, symbolized as h2, the eigen value and the total sum of squares are obtained and the results interpreted. For realistic results, we resort to the technique of rotation, because such rotations reveal different structures in the data. Finally, factor scores are obtained which help in explaining what the factors mean. They also facilitate comparison among groups of items as groups. With factor scores, one can also perform several other multivariate analyses such as multiple regression, cluster analysis, multiple discriminant analysis, etc.

26. Describe the several methods of factor analysis.

There are several methods of factor analysis, but they do not necessarily give same results. As such factor analysis is not a single unique method but a set of techniques. Important methods of factor analysis are:

1. the centroid method;
2. the principal components method;
3. the maximum likelihood method.

Before we describe these different methods of factor analysis, it seems appropriate that some basic terms relating to factor analysis be well understood.

1. **Factor:** A factor is an underlying dimension that account for several observed variables. There can be one or more factors, depending upon the nature of the study and the number of variables involved in it.
2. **Factor-loadings:** Factor-loadings are those values which explain how closely the variables are related to each one of the factors discovered. They are also known as factor-variable correlations. In fact, factor-loadings work as key to understanding what the factors mean. It is the absolute size (rather than the signs, plus or minus) of the loadings that is important in the interpretation of a factor.
3. **Communality (h2):** Communality, symbolized as h2, shows how much of each variable is accounted for by the underlying factor taken together. A high value of communality means that not much of the variable is left over after whatever the factors represent is taken into consideration. It is worked out in respect of each variable as under:
   h2 of the ith variable = (ith factor loading of factor A)2
   + (ith factor loading of factor B)2 + …
4. **Eigen value (or latent root):** When we take the sum of squared values of factor loadings relating to a factor, then such sum is referred to as Eigen Value or latent root. Eigen value indicates the relative importance of each factor in accounting for the particular set of variables being analysed.
5. **Total sum of squares:** When eigen values of all factors are totalled, the resulting value is termed as the total sum of squares. This value, when divided by the number of variables (involved in a study), results in an index that shows how the particular solution accounts for what all the variables taken together represent. If the variables are all very different from each other, this index will be low. If they fall into one or more highly redundant groups, and if the extracted factors account for all the groups, the index will then approach unity.
6. **Rotation:** Rotation, in the context of factor analysis, is something like staining a microscope slide. Just as different stains on it reveal different structures in the tissue, different rotations reveal different structures in the data. Though different rotations give results that appear to be entirely different, but from a statistical point of view, all results are taken as equal, none superior or inferior to others. However, from the standpoint of making sense of the results of factor analysis, one must select the right rotation. If the factors are independent orthogonal rotation is done and if the factors are correlated, an oblique rotation is made. Communality for each variables will remain undisturbed regardless of rotation but the eigen values will change as result of rotation.
7. **Factor scores:** Factor score represents the degree to which each respondent gets high scores on the group of items that load high on each factor. Factor scores can help explain what the factors mean. With such scores, several other multivariate analyses can be performed.We can now take up the important methods of factor analysis.

Centroid method of factor analysis in Research Methodology

The Centroid method of factor analysis, developed by L.L. Thurstone, was quite frequently used until about 1950 before the advent of large capacity high speed computers.* The centroid method tends to maximize the sum of loadings, disregarding signs; it is the method which extracts the largest sum of absolute loadings for each factor in turn. It is defined by linear combinations in which all weights are either + 1.0 or – 1.0. The main merit of this method is that it is relatively simple, can be easily understood and involves simpler computations. If one understands this method, it becomes easy to understand the mechanics involved in other methods of factor analysis. Various steps involved in this method are as follows:

1. This method starts with the computation of a matrix of correlations, R, wherein unities are place in the diagonal spaces. The product moment formula is used for working out the correlation coefficients.
2. If the correlation matrix so obtained happens to be positive manifold (i.e., disregarding the diagonal elements each variable has a large sum of positive correlations than of negative correlations), the centroid method requires that the weights for all variables be +1.0. In other words, the variables are not weighted; they are simply summed. But in case the correlation matrix is not a positive manifold, then reflections must be made before the first centroid factor is obtained.
3. The first centroid factor is determined as under:

   - The sum of the coefficients (including the diagonal unity) in each column of the correlation matrix is worked out.
   - Then the sum of these column sums (T) is obtained.
   - The sum of each column obtained as per (a) above is divided by the square root of T obtained in (b) above, resulting in what are called centroid loadings. This way each centroid loading (one loading for one variable) is computed. The full set of loadings so obtained constitute the first centroid factor (say A).
4. To obtain second centroid factor (say B), one must first obtain a matrix of residual coefficients. For this purpose, the loadings for the two variables on the first centroid factor are multiplied. This is done for all possible pairs of variables (in each diagonal space is the square of the particular factor loading). The resulting matrix of factor cross products may be named as Q1. Then Q1 is subtracted clement by element from the original matrix of correlation, R, and the result is the first matrix of residual coefficients, R1.* After obtaining R1, one must reflect some of the variables in it, meaning thereby that some of the variables are given negative signs in the sum [This is usually done by inspection. The aim in doing this should be to obtain a reflected matrix, R'1, which will have the highest possible sum of coefficients (T)]. For any variable which is so reflected, the signs of all coefficients in that column and row of the residual matrix are changed. When this is done, the matrix is named as ‘reflected matrix’ form which the loadings are obtained in the usual way (already explained in the context of first centroid factor), but the loadings of the variables which were reflected must be given negative signs. The full set of loadings so obtained constitutes the second centroid factor (say B). Thus loadings on the second centroid factor are obtained from R'1.
5. For subsequent factors (C, D, etc.) the same process outlined above is repeated. After the second centroid factor is obtained, cross products are computed forming, matrix, Q2. This is then subtracted from R1 (and not from R'1) resulting in R2. To obtain a third factor (C), one should operate on R2 in the same way as on R1. First, some of the variables would have to be reflected to maximize the sum of loadings, which would produce R'2. Loadings would be computed from R'2 as they were from R'1. Again, it would be necessary to give negative signs to the loadings of variables which were reflected which would result in third centroid factor (C).

Principal-components Method of Factor Analysis

Principal-components method (or simply P.C. method) of factor analysis, developed by H. Hotelling, seeks to maximize the sum of squared loadings of each factor extracted in turn. Accordingly PC factor explains more variance than would the loadings obtained from any other method of factoring. The aim of the principal components method is the construction out of a given set of variables Xj’s (j = 1, 2, …, k) of new variables (pi), called principal components which are linear combinations of the X_s

$$
\begin{aligned}
p_1 &= a_{11} X_1 + a_{12} X_2 + \dots + a_{1k} X_k \\
p_2 &= a_{21} X_1 + a_{22} X_2 + \dots + a_{2k} X_k \\
&\quad . \qquad\quad . \qquad\quad . \qquad\quad . \\
&\quad . \qquad\quad . \qquad\quad . \qquad\quad . \\
p_k &= a_{k1} X_1 + a_{k2} X_2 + \dots + a_{kk} X_k
\end{aligned}
$$

The method is being applied mostly by using standardized variables, i.e.,

$$\left(X_j - \overline{X}_j\right)^2/\sigma_j$$

The aij's are called loadings and are worked out in such a way that the extracted principal components satisfy two conditions: (i) principal components are uncorrelated (orthogonal) and (ii) the first principal component (p1) has the maximum variance, the second principal component (p2) has the next maximum variance and so on.

Following steps are usually involved in principal components method

1. Estimates of aij's are obtained with which X's are transformed into orthogonal variables i.e., the principal components. A decision is also taken with regard to the question: how many of the components to retain into the analysis?
2. We then proceed with the regression of Y on these principal components i.e.,
   $Y = \hat{y}_1 p_1 + \hat{y}_2 p_2 + \dots + \hat{y}_m p_m \ (m < k)$
3. From the $ aij and $yij , we may find bij of the original model, transferring back from the p's into the standardized X's.

Alternative method for finding the factor loadings is as under:

1. Correlation coefficients (by the product moment method) between the pairs of k variables are worked out and may be arranged in the form of a correlation matrix, R, as under:
   The main diagonal spaces include unities since such elements are self-correlations. The correlation matrix happens to be a symmetrical matrix.

2. Presuming the correlation matrix to be positive manifold (if this is not so, then reflections as mentioned in case of centroid method must be made), the first step is to obtain the sum of coefficients in each column, including the diagonal element. The vector of column sums is referred to as Ua1 and when Ua1 is normalized, we call it Va1. This is done by squaring and summing the column sums in Ua1 and then dividing each element in Ua1 by the square root of the sum of squares (which may be termed as normalizing factor). Then elements in Va1 are accumulatively multiplied by the first row of R to obtain the first element in a new vector Ua2. For instance, in multiplying Va1 by the first row of R, the first element in Va1 would be multiplied by the r11 value and this would be added to the product of the second element in Va1 multiplied by the r12 value, which would be added to the product of third element in Va1 multiplied by the r13 value, and so on for all the corresponding elements in Va1 and the first row of R. To obtain the second element of Ua2, the same process would be repeated i.e., the elements in Va1 are accumulatively multiplied by the 2nd row of R. The same process would be repeated for each row of R and the result would be a new vector Ua2. Then Ua2 would be normalized to obtain Va2. One would then compare Va1 and Va2. If they are nearly identical, then convergence is said to have occurred (If convergence does not occur, one should go on using these trial vectors again and again till convergence occurs). Suppose the convergence occurs when we work out Va8 in which case Va7 will be taken as Va (the characteristic vector) which can be converted into loadings on the first principal component when we multiply the said vector (i.e., each element of Va) by the square root of the number we obtain for normalizing Ua8.

3. To obtain factor B, one seeks solutions for Vb, and the actual factor loadings for second component factor, B. The same procedures are used as we had adopted for finding the first factor, except that one operates off the first residual matrix, R1 rather than the original correlation matrix R (We operate on R1 in just the same way as we did in case of centroid method stated earlier).

4. This very procedure is repeated over and over again to obtain the successive PC factors Other steps involved in factor analysis

- Next the question is: How many principal components to retain in a particular study? Various criteria for this purpose have been suggested, but one often used is Kaiser's criterion. According to this criterion only the principal components, having latent root greater than one, are considered as essential and should be retained.
- The principal components so extracted and retained are then rotated from their beginning position to enhance the interpretability of the factors.
- Communality, symbolized, h2, is then worked out which shows how much of each variable is accounted for by the underlying factors taken together. A high communality figure means that not much of the variable is left over after whatever the factors represent is taken into consideration. It is worked out in respect of each variable as under:
  h2 of the ith variable = (ith factor loading of factor A)2
  + (ith factor loading of factor B)2 + …
  Then follows the task of interpretation. The amount of variance explained (sum of squared loadings) by each PC factor is equal to the corresponding characteristic root. When these roots are divided by the number of variables, they show the characteristic roots as proportions of total variance explained.
- The variables are then regressed against each factor loading and the resulting regression coefficients are used to generate what are known as factor scores which are then used in further analysis and can also be used as inputs in several other multivariate analyses.

Maximum Likelihood (ML) Method of Factor Analysis

The ML method consists in obtaining sets of factor loadings successively in such a way that each, in turn, explains as much as possible of the population correlation matrix as estimated from the sample correlation matrix. If Rs stands for the correlation matrix actually obtained from the data in a sample, Rp stands for the correlation matrix that would be obtained if the entire population were tested, then the ML method seeks to extrapolate what is known from Rs in the best possible way to estimate Rp (but the PC method only maximizes the variance explained in Rs). Thus, the ML method is a statistical approach in which one maximizes some relationship between the sample of data and the population from which the sample was drawn.

The arithmetic underlying the ML method is relatively difficult in comparison to that involved in the PC method and as such is understandable when one has adequate grounding in calculus, higher algebra and matrix algebra in particular. Iterative approach is employed in ML method also to find each factor, but the iterative procedures have proved much more difficult than what we find in the case of PC method. Hence the ML method is generally not used for factor analysis in practice.

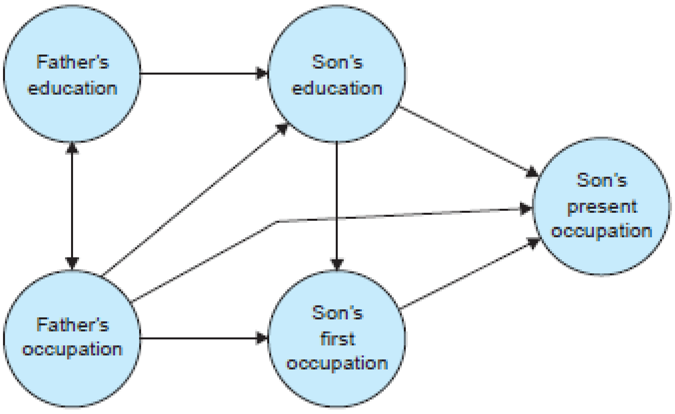
The loadings obtained on the first factor are employed in the usual way to obtain a matrix of the residual coefficients. A significance test is then applied to indicate whether it would be reasonable to extract a second factor. This goes on repeatedly in search of one factor after another. One stops factoring after the significance test fails to reject the null hypothesis for the residual matrix. The final product is a matrix of factor loadings. The ML factor loadings can be interpreted in a similar fashion as we have explained in case of the centroid or the PC method.

27. What is a Path Analysis in Research?

The term ‘path analysis’ was first introduced by the biologist Sewall Wright in 1934 in connection with decomposing the total correlation between any two variables in a causal system. The technique of path analysis is based on a series of multiple regression analyses with the added assumption of causal relationship between independent and dependent variables. This technique lays relatively heavier emphasis on the heuristic use of visual diagram, technically described as a path diagram. An illustrative path diagram showing interrelationships between Fathers’ education, Fathers’ occupation, Sons’ education, Sons’ first and Sons’ present occupation can be shown in the Fig. below

Path analysis makes use of standardized partial regression coefficients (known as beta weights) as effect coefficients. In linear additive effects are assumed, then through path analysis a simple set of equations can be built up showing how each variable depends on preceding variables. “The main principle of path analysis is that any correlation coefficient between two variables, or a gross or overall measure of empirical relationship can be decomposed into a series of parts: separate paths of influence leading through chronologically intermediate variable to which both the correlated variables have links.”

The merit of path analysis in comparison to correlational analysis is that it makes possible the assessment of the relative influence of each antecedent or explanatory variable on the consequent or criterion variables by first making explicit the assumptions underlying the causal connections and then by elucidating the indirect effect of the explanatory variables.

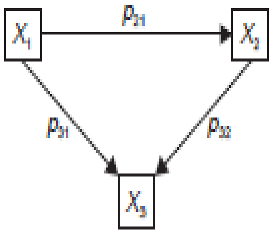


Path analysis makes

“The use of the path analysis technique requires the assumption that there are linear additive, a symmetric relationships among a set of variables which can be measured at least on a quasi-interval scale. Each dependent variable is regarded as determined by the variables preceding it in the path diagram, and a residual variable, defined as uncorrelated with the other variables, is postulated to account for the unexplained portion of the variance in the dependent variable. The determining variables are assumed for the analysis to be given (exogenous in the model).”

We may illustrate the path analysis technique in connection with a simple problem of testing a causal model with three explicit variables as shown in the following path diagram:

Path Diagram (with the variables)



The structural equation for the above can be written as:

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \begin{matrix} = \\ = \\ = \end{matrix} \begin{bmatrix} e_1 \\ p_{21}X_1 + e_2 \\ p_{31}X_2 + p_{32}X_2 + e_3 \end{bmatrix} = pX + e$$

where the X variables are measured as deviations from their respective means. p_{21} may be estimated from the simple regression of X_2 on X_1 i.e., $X_2 = b_{21}X_l$ and p_{31} and p_{32} may be estimated from the regression of X_3 on X_2 and X_1 as under:

$$\hat{X}_3 = b_{31.2}\, X_1 + b_{2.1}\, X_2$$

where $b_{31.2}$ means the standardized partial regression coefficient for predicting variable 3 from variable 1 when the effect of variable 2 is held constant.

In path analysis the beta coefficient indicates the direct effect of Xj (j = 1, 2, 3, ..., p) on the dependent variable. Squaring the direct effect yields the proportion of the variance in the dependent variable Y which is due to each of the p number of independent variables Xj (i = 1, 2, 3, ..., p). After calculating the direct effect, one may then obtain a summary measure of the total indirect effect of Xj on the dependent variable Y by subtracting from the zero correlation coefficient ryxj, the beta coefficient bj i.e.,

Indirect effect of X_j on Y = $c_{jy} = r_{yxj} - b_j$

for all j = 1, 2, ..., p.

Such indirect effects include the unanalysed effects and spurious relationships due to antecedent variables.

In the end, it may again be emphasised that the main virtue of path analysis lies in making explicit the assumptions underlying the causal connections and in elucidating the indirect effects due to antecedent variables of the given system.

Prepared by

Sidhartha Sankar Laha
Co-ordinator
Department of Economics
Cooch Behar Panchanan Barma University