ANALYSIS AND INTERPRETATION ON QUALITATIVE DATA

1.7.2 What is qualitative data?

Qualitative data can be found all around us. In the news papers we read, the television broad casts we view, memo received at work or the text messages we exchange via mobile phones, we come across a wealth of qualitative data every day. These naturally occurring sources of data are initially produced for a purpose other than our research; nevertheless, they still provides us with rich data to analyse. It is possible to carry out a highly informative research project which is based on the analysis of this type of qualitative data alone, and for some, such as media analysis, this is often the case. However, qualitative researchers often seek to generate their own data, which is gathered in the field. Methods such as interviews and observations enable researchers to collect rich data which is geared toward the research project at hand.

1.7.3. Natural occurring sources

Naturally occurring sources of qualitative data can have an advantage over primary data, since they are formed in the natural setting of social world, without any influence from the researcher. For example, we may be interested in seeking to understand exploitation and power relations in the workplace. The researchers could interview employees and directors within the organization and ask about experience of exploitation or feelings towards relationships with colleagues. However, this approach may only identify agencies at work which the interviews are aware of and are willing to discuss with the researcher. Observation techniques may result in a more accurate picture of the ways these issues manifest themselves. But there is still the potential for behaviour to be modified in the presence of the researcher. In instances such as this existing documentary sources such as internal memos, for contracts, and e-mail correspondence may provide the researcher with most valid data.

1.7.4 .Primary data

While qualitative data is fairly abundant in everyday life, researchers often have specific research questions in mind, for which they need to gather specific data. The use of interviews and observations techniques as methods of primary data collection is common in qualitative research. The notable advantage of primary data is precisely this ability of the researcher to be able to determine the content in which the data are

collected. Any research that concerns subject which do not present themselves readily in everyday life lends itself more to the collection of primary data. Since qualitative research is commonly associated with investigating meanings associated with subjects, or the ways in which people interpret their experiences, the researcher will often find she or he has to go out and get the data, rather than making use of what already exists.

When collecting data, it is important to be aware of the context in which the data was produced, since this will influence the ways in which analysis can be carried out.

Whether data are collected from existing documentary sources, or through data collection techniques devised explicitly for the project qualitative data can be categorized into textual, audio or visual information.

The example of qualitative data.

Type of data	Examples
Textual	Field notes, reflection journals, newspaper articles, memos, transcripts, e-mail / text message.
Audio	Audio recordings interviews, speeches, naturally occurring talk, radio broad casts, music.
Visual	Television, cinema, photographs, paintings, sculpture, video recording, focus groups or observations, video details.

Analysis of qualitative data means studying the organized material in order to discover inherent facts. These data are studied from as many angles as possible either to explore the new facts or to reinterpret already known existing facts. The content analysis, inductive analysis and logical analysis are mostly used in analysis of qualitative material.

.QUALITATIVE DATA ANALYSIS

The nature of the data depends mostly upon the type of tool or technique used by researchers for collecting these data. Most of information in behavioural sciences including social psychology and education is in the form of verbal and other symbolic behaviour. The verbal data gathered through questionnaires, observation or interview are

mostly qualitative in nature. These data indicate what people have said in their own words about their experiences and interactions in nature settings, and after careful analysis, the data provide useful and depth answers to the research questions of decision makers and information users. Patton (1982, p.22) emphasizes that:Qualitative data provide depth and detail. Depth and detail emerge through direct quotation and careful description. The extent of depth and detail will vary depending upon the nature and purpose of a particular study.

The responses to open-ended questions on a questionnaire are detailed and comprehensive. These responses are neither systematic nor standardized. However, they permit the researcher to understand situations as seen and felt by him. Since the responses to open-ended questions are longer and detailed, they help the researcher to understand in depth the points of view of other people, their level of emotion, their characteristics, their attitudes and values, and their experiences. Patton (1982, p.28) is of the view that responses to open-ended questions in the form of direct quotations reveal level of emotions of respondents, the way in which they have organized their world, their thoughts and experiences about certain happenings, and their perceptions.

The data gathered through participant observation or an open-ended/ unstructured interview are also descriptive in nature. These strategies are most comprehensive for understanding fully the complexities of a particular situation. Participant observation provides a first hand information to the researcher about some social even in depth and detail. Data gathered through participant observation generally include (i) description of the setting of the social situation; (ii) activities that take place in the setting; and (iii) description about people who participated in the activities and their extrinsic behaviour during the activities. The descriptions may be in the form of field notes specifying some basic information pertaining to the place where the observation takes place, the persons present during the observation, nature of the settings, type and nature of various types of interactions and activities during the observation. The field notes taken during observation contain direct quotations of the people who participated in the observation as well as the observer's own feelings and reactions.

It is not possible to find out what is in and on other individuals' mind while observing their extrinsic behaviour. Through participant observation, it is difficult for an

observer to know the feelings, thoughts and intentions of others and also about the behaviours that took place in earlier situations. However, through open-ended/ unstructured interview, it is possible to find out from people those things which had happened earlier or could not be observed during the participant observation. It provides framework within which the researcher should be able to gather information from people conveniently and accurately. The information mostly pertains to a programme, the reaction of participants about the programme and the type of change the participants perceive in themselves after their involvement in the programme. The data are mostly in the form of responses to structured and unstructured questions put to the respondents by the researcher during an informal conversation. The responses are generally direct quotations from respondents in their own words and provide details about situations, events, people, experiences, behaviours, values, customs, etc. The information gathered during or after an interview includes notes taken by an interviewer alongwith his detailed comments about what people say about their experiences, what they think and feel about the phenomena under study, and what they know about the phenomena.

.CONTENT ANALYSIS

Content analysis is concerned with the classification, organization and comparison of the content of document or communication. In the context of communication research. Berelson (1952) remarks that:

Content analysis is a research technique for the objective, systematic, and quantitative description of the manifest content of communication.

Cartwright (1970) uses the terms “content analysis” and “coding” interchangeably as both the processes involve objective, systematic, and quantitative description of any symbolic behavoiur. Since content analysis is concerned with the classification, evaluation and comparison of the content of communication or document, it is sometimes referred to as “documentary activity” or “information analysis”. The communication may be in the form of responses to open-ended questionnaire, conversation as a result of interview, or description of an observed activity. It may also be in the form of official records (census, birth, accident, crime, school, institutional and personal records), judicial decisions, laws, budget and financial records, cumulative records, courses of study,

content of text books, reference works, newspapers, periodicals or journals, prospectus of various educational institutions or universities, etc., direct quotations, and notes of an interview.

Berelson (1952) has specified three broad approaches that a researcher may adopt in content analysis. These include: (i) characteristics of content; (ii) producers or causes of content; and (iii) audience or effects of content. In any single study, he may or may not adopt more than one of these approaches.

Characteristics of Content

In this approach, the researcher is interested primarily in the characteristics of the content itself. He may focus either on the substantive nature of the content or upon the form of content. Berelson (1952), as quoted by Cartwright (1970), has listed six uses which are concerned primarily with substantive characteristics of the content. In the first two of these, the researcher either tries to describe trends in communication content over periods of time by employing methods for sampling the total flow of communication at successive points in time and to use the same system of classification throughout or he attempts to trace the development of scholarship in the publication of reputed scientific journals.

In the next two uses, the researcher attempts to compare the content materials coming from different sources. He may be interested either to disclose international differences in communication content or to compare media or "levels" of communication. Mc Granahan and Wayne (1948) compared the major themes of the most popular dramas appearing in Germany and America in the year 1927 and 1910. Cartwright (19700 points out that substantially similar differences were found between the two countries. He has reported that other studies comparing different countries have been made in terms of such media as radio, newspapers, and text books, and a few comparable interviewing surveys have also been conducted in different countries. It may be pointed out that in making such cross-national comparisons, the researcher may face problems of sampling and of translations. But such types of studies have produced useful comparative data for understanding national differences.

In the fifth use the researcher may be interested in studying the role of various mass media in moulding public opinion. For example, differences in partisanship among

newspapers, journals, radio and T.V in the role of a certain voluntary organization in the National Literacy Mission Programmes may be studied by some researcher to identify which media favour more the role of the organization in comparison to other.

In the sixth use of the analysis of characteristics of systematic material, the observed substance of communication content is evaluated against standards adopted by the researcher. Studies are conducted to evaluate the social contributions of the media of communication. Such an evaluation can be made by comparing actual performance against some specified explicit and precise standards such as "fairness", "objectivity" or "balance".

Producers or Causes of Content

In the second major approach, the researcher attempts to draw valid inferences about the nature of the procedure of the content or the causes of the symbolic material from the characteristics of the material itself. In some situations where the researcher cannot study the producer of the content (communicator) directly, but has access to the communicated material (content), this approach is used to make valid inferences about the nature of the producer of the content. In other situations, where a researcher can persuade an individual to produce symbolic behaviour as a response to standard conditions, the characteristics of such behaviour are generally taken as a very acceptable indication of the individual's own characteristics. Cartwright (1970) has highlighted four uses of content analysis which illustrate various ways in which researchers have tried to construct a picture of the communicator from his symbolic materials or products. The first use in confined to a number of studies in which intentions and attitudes of communicators have been inferred from an analysis of their communications. As illustration of this type of content analysis is the study of Leites et al. (1951) of speeches delivered in 1949 by members of Soviet elite in celebration of Stalin's birthday. These speeches were analyzed so as to know the attitudes of the speakers towards Stalin. It was found that sharp differences existed in the image of Stalin as revealed by the old Bolsheviks and by other speakers.

Assessment of psychological state of persons or groups using contents of clinical interviews, protective tests, biographies, autobiographies, diaries, letters and other personal documents illustrates second important use of content analysis. For example,

Baldwin (1942) studied the motivational and ideational clusters in a writer's personality after recording frequencies with which certain themes were contiguous to another within a sample of letters written by him. To detect the existence of propaganda or to obtain political and military intelligence illustrate other two uses of content analysis. For example, during the World War II the United States Department of Justice, making use of content analysis developed by Lasswell (1949), revealed remarkable similarities between allegedly native-fascist propaganda and the propaganda of Nazis. During war time, there is need for intelligence (political or military) about the activities of hostile nations. For this, methods using content analysis of various communication material, have been developed.

Audience and Effects of Content

In the third major approach to content analysis, the researcher interprets the content so as to reveal something about the nature of its audience or its effects. He takes the content material as a basis for drawing inference about the characteristics of the audience for whom the material (content) is designed, or about the effects of communication. This approach is illustrated by the following three uses of content analysis.

Some studies of the content of mass media assume that the material communicated through these media reflect the feeling or thought of the group or population existed at that time. For example, Hart (1933) analyzed the content of popular magazines in the United States over a period of years from 1900 to 1930. He found what he took to be evidence for a decline in the status of religion and an increase in tolerance of sexual freedom during this period.

In the second use of content analysis, it is assumed that there is a more or less approximate correspondence between the content of the mass media and the cognitive behaviour of the audiences exposed to the media. If a particular theme is stressed in the media at some place and time, it is assumed that this theme will be prominent in the thinking of the audience or population. A researcher, for example, may study that day time TV programmes deal predominantly with the themes concerning personal problems rather than public affairs.

Studies dealing with the description of attitudinal and behaviour responses to communications illustrate the third use of content analysis. Berelson (1942) points out three ways in which content analysis has been used to study the effects of communications. The first consists in analyzing materials which were produced in response to some specific communication. A researcher, for example, may study published reactions to a book on a theme highlighting India Culture. In the second kind of analysis, a researcher may attempt to show empirical relations between the content of a communication and responses to it. To illustrate, a researcher may attempt to trace a relationship between the frequency of arguments concerning "Privatisation of Higher Education" in the various media and the number of university teachers who could recognize the argument. The third kind of analysis attempts to make a direct inference to

the effect of content without any reference to response data themselves. Using this type of approach, Lasswell and Blumenstock (1939) analyzed the themes employed by communists in Chicago in 1930's and concluded that their propaganda was relatively ineffective because it was not in line with fundamental values of citizens to whom it was addressed.

STEPS IN CONTENT ANALYSIS

The preceding discussion was confined to the details about the objectives of content analysis and its uses. The following section describes the steps involved in the process alongwith some issues relating to this operation.

Defining the Unit of Analysis

The unit (material) may be confined to single words, to phrases, to complete sentences, to paragraphs, or to even larger amount of material such as articles or to complete books. Either of these can be considered as an entity whose specified characteristics can be determined and analyzed. Hayman (1968) suggests that the unit should be comprehensive enough to provide meaning through some content atleast, but small enough not to allow subjectivity in its use.

Specifying Variables and Categories

Once the unit is defined, the researcher conducts its analysis so as to create reproducible or objective data for scientific treatment and generalization beyond the specific set of symbolic material analyzed. For converting symbolic material into objective data, it is necessary to specify the "variables" explicitly in terms of which descriptions are t be made. The variables are sometimes referred to as "dimensions" or "types of attributes". A few examples of such variables are: number of words, percentages of personal pronouns, attitude towards privatization, attractive traits of teachers, degree of confidence in a friend, etc. After the selection of some variable, viz., degree of confidence in a friend, there are many ways in which this variable may be broken down into categories as: (1) Unqualified balanced, (2) Qualified confidence, (3) Confidence and mistrust equally balanced, (4) Qualified mistrust, (5) Unqualified mistrust, (6) Question not asked by interviewer, (7) Question asked, but answer not classifiable in above categories. A second classification of categories of the same variable may be: (1) High, (2) Low, (3) Not classifiable in either. It may be pointed out that if two

independent persons were to code the same material, one using the first set of categories and the other using the second, they would come out with different descriptions of the same material. Hence, explicit specification of the system of categories used with each variable is necessary for reproducible analysis.

There is need for framing explicit rules specifying what features of the content are to be taken as indication that it falls in one category rather than another. A statement of these rules constitute the operational definition of the category. These specific rules are helpful in arriving at an agreed system of coding if the analysis is conduced by two or more independent analysis.

In developing such an operational definition, an analyst may begin by designating the units of analysis that are to be used. In this context, Cartwright (1970) suggests that there are basically two kinds of units to be specified. The first of these he calls as the "recording unit" and the second as the "content unit". The recording unit is the specific segment of the content that is characterized by placing it in a given category. The content unit is the largest body of the content that may be examined in characterizing a recording unit. For example, in the coding of free-answer interviews, the answer to a single question is often taken as the recording unit and whole set of related questions as the content unit.

Another aspect of the operational definition of a category consists in specifying the 'indicators' which determine whether any given unit should fall within the category. For example, while considering the category "degree of confidence in a friend", the statement "stood by me and my brother to protect our family from armed bandits" can be taken as an indicator. If an analyst is coding a large content or a very large number of interviews, he would encounter many statements or indicators for a particular category. Thus, a category consists of a range of possible indicators, all of which are given the same label and are therefore treated equivalently in all the subsequent analysis.

Frequency, Direction and Intensity

Once the unit is defined and the variables alongwith their categories to be employed specified, the analyst will classify units, in the material to be analyzed, according to frequency, direction and intensity.

For frequency, the analyst merely counts the number of units which fall into each of his categories. Cartwright (1970, p.440) refers to it as "unit of renumeration". The "unit of renumeration" and "recording unit" are not necessarily the same. But when the analyst merely counts the number of recording units which get a certain categorization, the recording unit is exactly the same as the enumeration unit. For example, in the analysis of public speech by an economist, the number of times "privatization of higher education" may be employed as an "argument" for a certain policy of government. In this case an "argument" is taken both as the recording unit and the enumeration unit.

In another example, suppose an analyst analyses an editorial on "Privatisation of Higher Education" for its favourableness or unfavourableness. For purposes of quantification, he counts the number of column inches of the whole editorial. In this case, a column inch would be the unit of renumeration, whereas the editorial as a whole would be the recording unit and hence the two units are not identical.

In certain situations, it is useful to further classify the units according to direction and intensity. Direction refers to whether the reference was favourable, unfavourable, or neutral. It might be pleasant-unpleasant, interesting-uninteresting, threatening-non-threatening. Intensity indicates the emotional impact of the units analyzed. Is it large or small, and in what direction?

Judging direction and intensity is more subjective than merely counting for frequency.

Contingency Analysis

The contingency analysis aims at considering the content within which the unit is found. A researcher should consider the favourableness or unfavourableness of a single unit in the light of the remainder of the communication so that its real meaning might not be lost.

Sampling

One of the major and practical problems in content analysis is sampling. The unit which a researcher analyses must be representative of the total material with which he is concerned, so that the results can be generalized. Invariably a researcher undertakes the analysis of a specific content in order to reveal something about the universe of data than just those symbolic material with which he deals.

In any give study, the universe of symbolic materials, that should be selected will depend upon the purposes of the investigator. If, for example, the objective is to compare the editorial content of "The times of India" against some established standard or norm, the universe under consideration should be all editorial content appearing in all issues of "The Times of India" over a certain period of time. After the specification of the universe, proper procedures for drawing a representative sample of that universe must be employed.

CONSTRUCTING THE CONTENT ANALYSIS OUTLINE

Cartwright (1970) has suggested the following six steps for arriving at a satisfactory content analysis outline.

Step 1. Specify Needed Data

In laying out a satisfactory analysis outline, a researcher should clearly specify the data that are required by him in the total research design so that he may face less difficulties in the long run. The specification of needed data is helpful in planning the final tables which the researcher may use at later stages of content analysis.

Step 2. Map out Plans for Tabulation

A researcher can avoid a number of problems if he makes explicit plans for the tabulation of coded data. He should decide in advance whether the coded data are to be punched on cards for machine processing or to be tabulated by hand.

Step 3. Lay out the Skeleton of the outline

The researcher should list the variables in terms of which the content is to be coded. For example, if the study pertains to analyzing interviews, these variables will be used to classify not only various features of the answers to questions about the psychological make up of the respondent but also such matters as his age, income, marital status, and other demographic and behavioural characteristics. In listing the variables to be included in the outline, the researcher should ensure that all information needed on the punch cards is placed on some variable. The outline should contain provision for coding the name of the study, the number of each enumeration unit (interview, issue of journal or newspaper, etc), the name of coder, and any other relevant information.

The researcher may make use of the following list of variables summarized by Berelson (1952) under two broad headings of "what is said" and "How it is said".

(i) What is said

1. Subject matter: what is the communication about?
2. Direction: Is the treatment favourable or unfavourable towards the subject?
3. Standard: What is the basis on which the classification of direction is made?
4. Values: What goals are explicitly or implicitly revealed?
5. Methods: What means or actions are employed to realize goals?
6. Traits: What characteristics of persons are revealed?
7. Actor: Who initiates actions?
8. Authority: In whose name are statements made?
9. Origin: What is the place of origin of the communications?
10. Target: To whom is the communication particularly directed?

(ii) How it is said.

1. Form of communication: Is it fiction, news, television, etc?
2. Form of statement: What is the grammatical or syntactical form of the unit of analysis?
3. Intensity: How much strength or excitement value does the communication have?
4. Device: What is the rhetorical or Propagandistic Character of the Communication?

Step 4. Fill in Categories for Each Variable

A researcher should use a system which is exhaustive with mutually exclusive categories. The system is exhaustive if there is a category in which to place every relevant item which may be found in the content. Its categories are mutually exclusive if there is one and only one place to put an item within that system of categories. After defining all the categories in a system, a manual of instructions should be prepared with operational definitions of the categories.

Step 5. Establish Procedure for Unitizing Material

It is essential for a researcher to establish a procedure for unitizing material. Cartwright (1970, p.458) is of the opinion specific working definitions to be used in the content analysis should be formulated in such a way that various coders can all unitize

the same material in the same way. These definitions should be written down as a part of the coding instructions.

Step 6. (i) Tryout the Analysis Outline and Unitizing Procedure

The analysis outline and unitizing procedure on a sample of material should be given a tryout in order to discover what modifications are needed. This trying out of the coding procedures is also used as a training for those who are to do the final coding.

(ii) Selection and Training of Coders

After constructing the analysis outline, the researcher should select coders who are able to use the analysis outline as intended and in a standardized manner. It is assumed that several coders are involved in actual coding. The number of coders who are able to use the analysis outline in a standardized manner will depend upon the volume of material (content). If the volume of material to be analyzed is large, several coder will do the actual coding. In case the volume of the content is too small, the researcher may act as the coder himself. However, in such situations it is desirable to have an independent coder also so that the whole procedure can be objectified. Whatever the case, certain skills and abilities are essential for satisfactory coding.

The successful and meaningful use of a well-developed outline depends upon the selection of efficient coders and their effective training in the outline being used under good supervision so that the proper procedures of coding are followed. If the content to be analyzed is large, the process of coding involves the repetitive application of the analysis outline to the material. It demands same operational definition of categories, the same frame or reference, the same degrees of differentiation, etc. throughout the entire coding operation. In such cases, a person who is easily satiated with repetitive work should not be selected as a coder.

After the selection of efficient coders, it is necessary to train them in the use of analysis outline so that they have a full understanding of the objectives of the project.

Step 7. Mechanics of Coding

A good analysis of material requires that regularized procedures be established for their storage, assignment to coders, and recording. It is convenient to package together a collection of material and get the package coded. When the coded package is returned to a central storage place, a new package is taken. In the distribution of material to coders, it

is desirable to randomize the material so as to eliminate any systematic biases among the coders. This will also ensure that all the material gets coded once and only once. The same precautions should be taken in collecting and storing the sheets upon which coding has been recorded.

At the stage of final coding, the coder can add new categories to some of the variables of the outline if he comes across a recording unit for which there is no category. However, the merit of adding a new category should be assessed by determining whether the new category would be meaningful within the rationale of the system of categories. It is also desirable to hold periodic discussion among the coders to ensure that same frame of reference and operational definitions of categories are maintained throughout the coding period.

Cartwright (1970) has suggested the procedure of "Check Coding" for assessing the reliability of coding. In this procedure, a certain percentage of the content is independently recorded by a "Check Coder" who is taken as a sort of "Criterion". The records of disagreements as a result of check coding should be maintained and tabulated in various ways. The records should be tabulated separately for each coder and the interpretation of differences among coders should be taken into consideration in the final results.

INDUCTIVE ANALYSIS

Inductive analysis means that patterns, themes, and categories of analysis emerge out of the data. In this analysis, researcher looks for natural variation in the data.

Patton describes these approaches as: (i) Indigenous typologies, and (ii) Analyst-constructed typologies. "Indigenous typologies" approach is called as the "emic" approach to analysis in anthropology. In this approach, cultural behaviour is studied in terms of the inside view of human events rather than imposed from cross- cultural classification of behaviour. In order to understand the thoughts and views of a particular group of people, it is necessary that the whole analysis of experience must be based on their concepts. Hence, indigenous typologies approach require an analysis of the verbal categories used by the sample group or participants of the programme so that complexity of reality is broken into meaningful parts. The 'verbal categories' used by the participants

or others who are associated with a particular programme are ‘labelled’ and given names so that one can be separated from others. Once these labels have been identified from an analysis of what participants in the programme have said, the next step is to identify and list the characteristics or attributes that distinguish one from the other. At a later stage, these categories become “themes” which are important throughout data study analysis and in the final report.

Suppose a study aims at enhancing the achievement level of slow learners studying in eighth grade of a high school. In observation and interviews at the selected high school, it is necessary to understand the ways in which teachers categorize students with regard to problems of low achievement in unit tests, absenteeism, and irregularity in the completion of home assignments. The teachers may describe the extremes as the ones who fail in almost all the unit tests and never complete their home assignments. Another teacher may mark them as irregular in almost all teaching periods and in class tests. The borderlines, on the other hand, may be described as the ones who fail in more than half of the unit tests. Another teacher may characterize them as different from extremes but their performance in the unit tests being far below than that of other students.

Not all teachers use the same criteria to distinguish “extremes” from “borderlines”, but all of them use the same labels in talking about them. In order to study the impact of any remedial programme activities on the achievement of slow learners, it is important to understand the differences in programmes between the “extremes” and “borderlines”. Moreover, it is not possible to analyze fully the situations in the programme as understood by teachers and experienced by students without understanding the indigenous typology of “extremes” and “border-lines”. This typology will have immense implications for how the programme activities were organized and the extent to which different strategies are to be developed to deal with the “extremes” and “borderlines”. Finally, these categories become themes which are important throughout the data analysis and in the final report.

In the “analyst-constructed typologies” approach, the researcher looks for patterns, categories and themes for which the participants of the programme do not have labels or terms, and the researcher himself constructs typologies to explain variations and contrasts in activities, participants and others associated with the programme. However,

the researcher must take utmost care that such constructed typologies are accurate and make sense when presented to the people. The main objective of these typologies is to make descriptions based on an analysis of the patterns that appear in the data so as to make interpretations about the nature of the progrmme.

The patterns in qualitative data are converted into meaningful categories. Guba (1978), quoted by Patton (1982, pp.311-312), suggests several steps for converting field notes and observations about issues and problems into systematic categories of analysis using the principle of "convergence" i.e. "figuring out what things fit together" and "divergence". Categories are further judged by two criteria: "Internal homogeneity" and "external homogeneity". The first criterion concerns the extent to which the data that belong to a certain category hold together. The second criterion concerns the extent to which differences among categories are significant and clear. The existence of a large number of unassignable or overlapping data items indicates that there are some basic faults in the category system. Using these two criteria, the researcher develops several different classification systems of categories and then establishes some priorities to determine which category systems are more important than others.

By "divergence" Guba means that the researcher must deal with how to "flush out" the categories. He suggests that this can be done by process of extension (building on items of information already known), bridging (making connections among different items), and surfacing (proposing new information that ought to fit and then verifying its existence).

The procedures suggested by Guba for analyzing qualitative data are flexible. Since identification and naming patterns, themes, and categories is a creative process, the researcher must rely on his own intelligence, experience and judgement.

LOGICAL ANALYSIS

Indicative analysis is used for representing patterns as dimensions or categories, either using participant-generated constructions or evaluator-generated constructions. It is sometimes useful to cross-classify different dimensions to generate new insights about how the data can be organized and to look for patterns that may not have been recognized in the initial inductive analysis. Logical analysis aims at creating potential categories by

crossing one typology with another, and then moving back and forth between the logical construction and the actual data for creating a "new typology" using cross- classification matrices.

According to Patton (1982),Creating cross-classification matrices is an exercise in logic. This procedure involves creating potential categories by crossing one dimension or typology with another, and then working back and forth between the data and one's logical constructions, filling in the resulting matrix. This logical system will create a new typology all parts of which may or may not actually be represented by the data. Thus, the analyst moves back and forth between the logical construction and the actual data in ongoing search for understanding through description.

The researcher must be extremely careful in using this kind of analysis. He should be sensitive to interpreting the possibility of a category of activity or behaviour that has either been overlooked in the data or that is logically a possibility in the setting but has not been manifested.

VALIDATION OF QUALITATIVE ANALYSIS

This section is concerned with the major strategies that are helpful for validating and verifying the results of qualitative analysis. Patton (1982, 99.327-334) has listed the following seven major strategies for validation of results:

Rival Explanations

Once the researcher after qualitative analysis has described the patterns and their explanations, it is important to look for rival or competing themes and explanations both inductively and logically. Inductively it implies looking for other ways of organizing the data that might lead to different results. Logically it involves searching for other logical possibilities and then finding if those possibilities can be supported by the data. However, it may be noted that when considering rival hypotheses and competing explanations, the strategy to be employed by the researcher is not one of attempting to disprove the alternatives, but to look for data that support alternative explanations. In this strategy, the researcher should give due weightage to supporting evidence and look for the best "fit" between data and analysis.

Negative Cases

The search for negative cases and instances that do not fit within the identified pattern and their understanding is also important in the verification and validation of results.

Triangulation: Reconciling Qualitative and Quantitative Data

This type of triangulation aims at comparing data collected through some king of quantitative methods with data collected through same kind of qualitative methods. It is highly likely that qualitative methods and quantitative methods will eventually lead to different findings and not to a single and well integrated picture of the situation. It is because qualitatively data are commonly used for "generating hypotheses" or "describing hypotheses" and quantitative data are used to "analyze outcomes", or "verify hypotheses". However, in endorsing the notion triangulation, Trend (1978), quoted by Patton (1980, p.330), maintains that it is useful to bring a variety of data and methods to bear on the same problem in order to reduce system bias in interpreting results of study. The findings of some studies could be strengthened by supplementing qualitative approach with quantitative analysis.

Triangulation: Comparing Multiple Qualitative Data Sources

This type of triangulation involves comparing and cross checking consistency of data derived by different means at different times using qualitative methods. It means (i) Comparing observational data with interview data; (ii) Comparing observational data with questionnaire data; (iii) Comparing what participants of a programme say in public with what they say in private; (iv) Checking for the consistency the opinion of the participants about a programme over a period of time and (v) Comparing the opinion of the participants of a programme with others who were associated with programme in one capacity or the other. The triangulation of data sources within qualitative methods will seldom lead to a single totally consistent picture. But such triangulation is helpful to study and understand when and why there are differences.

Triangulation: Multiple Perspectives from Multiple Observers

The aim of this kind of triangulation is to involve triangulating observers or using several interviews so as to reduce the potential bias or subjectivity as a result of observation by single observer.

Design Checks: Keeping Methods and Data in Context

The nature of research design and methodology also contribute to distortion in results. Sampling gives rise to three type of errors. The errors may be due to: (i) distortion in the situations that were sampled for observation; (ii) distortion introduced by the time periods during which observations took place; (iii) distortion because of selectivity in the people who were sampled either for observation or interviews. Thus the researcher must be careful to limit results of his study to those situations, time periods, people and contents for which the data are applicable.

Evaluator Effects

The presence of researcher during the observation or interview can distort the results of study. The distortion may be due to: (i) reactions of programme participants and others associated with it to the presence of researcher; (ii) changes in the researcher during the process of observation or interview; (iii) biases of researcher, and (iv) incompetence of the researcher. The presence of a researcher during observation or interview may create a halo effect and consequently the participants of the programme are motivated to "show off". Their deviation from the normal behaviour will lead to distorted findings. It is desirable to undertake long term observations for minimizing the halo effect. Researcher sometimes becomes personally involved with programme participants and therefore lose their sensitivity to the full range of events occurring during the process of observation of interview. A record of the changes in the researcher, field notes and conversation with the people associated with the programme are helpful to overcome evaluator effects.

CORRELATION

Introduction

The method of correlation is developed by Francis Galton in 1885, as he published a paper on the topic 'Regression towards mediocrity in Hereditary stature'. Karl Pearson extended Galton's concept of regression and evolved the methods of correlation which is used widely today.

Meaning and Definition of Correlation

The most widely used measure of correlation is the pearson's product moment correlation is the pearson's product moment correlation coefficient (r). This measure is used where the variables are quantitative ie of the interval or ratio scale. Other methods of correlation have been developed for the use with nominal and ordinal variables.

Definition of Correlation

"Whenever two variables of the same group are so related that the increase or decreases are correspond to the increase or decrease to another or conversely, increase or decrease corresponds to the decrease or increase to another they are said to be correlated".

The thing, which indicates how change in one variable affects the other variable is called correlation.

Correlation is the study of relationship between variables. (eg) Relationship between Intelligence Quotient and Academic Achievement.

Types of Correlation

1. Positive correlation
2. Perfect positive correlation
3. Zero correlation
4. Negative correlation
5. Perfect Negative Correlation.

Positive correlation

If the value of one variable increases or decreases then the value of another variable will increase or decrease then it is called positive correlation.

(eg) As intelligence increases the academic achievement increases.

Negative correlation

If the value of one variable increases then value of another variable decreases, in the same way if the value of one variable decreases then the value of another variable increases it is called Negative correlation.

(eg) As practice time in English increases, the errors in writing English decreases.

Zero correlation

If there is no relationship between the variables it is called zero correlation

(eg) There is no relationship between the colour and intelligence of a person.

4.13.1. Methods of co-efficient of correlation

1. Rank difference method
2. Product moment correlation]
3. Partial correlation
4. Multiple correlation
5. Bisevial correlation
6. Point Bisevial correlation
7. Tetrachoric correlation
8. Phi-correlation.

Methods of Finding Correlation Co-Efficient

4.13.2. Formula for the Spearman's Rank Correlation C0-efficient)

$$\ell = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)}$$

D = Difference in Rank

N = No of Pairs

4.13.3. Steps for calculating rank correlation co-efficient

1. Find Ranks for the data X
2. Find Ranks for the data Y
3. Find Difference in Ranks
4. Find the square for the differences
5. Apply the formula (4.13.2) for calculating Rank correlation Co-efficient

Find Rank Correlation co-efficient for the data

Marks in Maths (x)	Marks in science ®
50	45
65	52
72	75
82	85
32	54
46	48

Calculation of Rank Correlation Co-Efficient

Marks in Maths (x)	Marks in science ®	R_1	R_2	D (R_1-R_2)	D^2
50	45	4	6	2	4
65	52	3	4	1	1
72	75	2	2	0	0
82	85	1	1	0	0
32	54	6	3	3	9
46	48	5	5	0	0
					$\Sigma D^2 = 14$

$$\rho = 1 - \frac{6\sum D^2}{N(N2 - 1)}$$

- **ρ** – co-efficient of correlation
- **D** – Difference between ranks
- $\sum \mathbf{D^2}$ - Sum of squares of difference between ranks
- **N** – Number of pairs

$$=1-\frac{6\times14}{6(6^2-1)}$$

$$=1-\frac{84}{6\times35}=1-\frac{84}{210}=1-0.4=0.6$$

$\boldsymbol{\rho}$ = 0.6

If $\boldsymbol{\rho}$ > 0.5 High

If $\boldsymbol{\rho}$ = 0.5 Moderate

If $\boldsymbol{\rho}$ < 0.5 Low

Find rank correlation co-efficient for the data

X	**Y**
10	12
12	18
18	25
18	25
15	50
40	25

Calculation of Rank Correlation Co-Efficient

X	**Y**	**R_1**	**R_2**	**D**	**D^2**
10	12	6	6	0	0.00
12	18	5	5	0	0.0
18	25	2.5	3	0.5	0.25
18	25	2.5	3	0.5	0.25
15	50	4	1	3	9
40	25	1	3	2	4
					ΣD^2 =13.50

$$\ell=1-\frac{6\times13.5}{6(6^2-1)}$$

$$=1-\frac{81}{6\times35}$$

$= 1 - \frac{81}{210}$

=1-0.39 = 0.61

= 0.61

Very high positive correlation.

Find rank correlation co-efficient for the data

Case	X	Y
A	82	74
B	75	69
C	90	70
D	63	64
E	57	71
F	85	79
G	82	80
H	65	68

Calculation of Rank Correlation Co-Efficient

Case	X	Y	R_1	R_2	D	D^2
A	82	74	3.5	3	0.5	1.25
B	75	69	5	6	1	11
C	90	70	1	5	4	16
D	63	64	7	8	1	10
E	57	71	8	4	4	16

F	85	79	2	2	0	0
G	82	80	3.5	1	2.5	6.25
H	65	68	6	7	1	1
						ΣD^2 =41.5

$$\ell = 1 - \frac{6 \times 41.5}{8(8^2 - 1)}$$

$$= 1 - \frac{249}{8 \times 63}$$

$$= 1 - \frac{249}{504}$$

= 1-0.49 = 0.51

ℓ = 0.51

Find rank correlation co-efficient for the data

X	**Y**
10	16
15	16
11	24
14	18
16	22
20	24
10	14
8	10
7	12
9	14

Calculation of Rank Correlation Co-Efficient

X	**Y**	**R_1**	**R_2**	**D**	**D^2**
10	16	6.5	5.5	1	1.00
15	16	3	5.5	2.5	6.25
11	24	5	1.5	3.5	12.25

14	18	4	4	0	0.0
16	22	2	3	1	1.00
20	24	1	1.5	0.5	0.25
10	14	6.5	7.5	1	1.00
8	10	9	10	1	1.00
7	12	10	9	1	1.00
9	14	8	7.5	0.5	0.25
					ΣD^2 =24

$$\ell = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)}$$

$$\ell = 1 - \frac{6 \times 24}{10(10^2 - 1)}$$

$$= 1 - \frac{144}{10 \times 99} = 1 - \frac{144}{990}$$

= 1-0.15 = 0.85

ℓ = 0.85

ρ

PRODUCT MOMENT CORRELATION

Pearson product moment correlation

Pearson product moment correlation coefficient is widely used in research and measurement. There are several forms of correlation co-efficient, but it is most popular and extensively used technique. This method is an advanced and accurate method. It is however complicated then rank difference method.

The most important disadvantage of the rank difference method is that it is useful only if the number of scores is less. If the value of N is more than 30, then it is not useful. The method becomes complicated as the number of scores increases. The value of co-efficient of correlation calculated by using the rank difference method is not supposed to be technically very much accurate. It is therefore, better to use the product moment method.

In Pearson's method, various formulae are applied for calculation of here, however, we shall use the simplest formula

4.14.1. Formula for calculating Product Moment Correlation

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

Were: r = Pearson's Co-efficient of correlation

X = deviation of X from assumed mean.

Y= deviation of Y from assumed mean.

Σ = Sum

N= Size of sample

Where: X and Y are the original scores.

The application of the formula for computing the correlation co-efficient has been illustrated by the following example.

4.14.2 Steps for calculating Product Moment Correlation Co-efficient

1. Find Sum for X
2. Find Sum for Y
3. Find Sum for $\mathbf{X^2}$
4. Find Sum for $\mathbf{Y^2}$
5. Find the Sum of the product of X and Y
6. Apply the formula (4.14.2)

Calculate the product moment co-efficient of correlation from the given data using raw scores

X	**Y**
1	4
4	3
6	2
5	7
2	1
1	5
6	10

2	5
8	11
5	12

Calculation of Product Moment Correlation Co-Efficient

X	Y	X^2	Y^2	XY
1	4	4	1	16
4	3	12	16	9
6	2	12	36	4
5	7	35	25	49
2	1	2	4	1
1	5	5	1	25
6	10	60	36	100
2	5	10	4	25
8	11	88	64	121
5	12	60	25	144

N=10 $\Sigma X = 40$ $\Sigma Y = 60$ $\Sigma XY = 288$ $\Sigma X^2 = 212$ $\Sigma Y^2 = 494$

$$r = \frac{N\Sigma XY - (\Sigma X)(Y)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

$\Sigma X = 40$

$$r = \frac{10 \times 288 - (40)(60)}{\sqrt{10 \times 212 - (40)^2}\sqrt{10 \times 494 - (60)^2}}$$

$\Sigma Y = 60$

$$r = \frac{2880 - 2400}{\sqrt{2120 - 1600}\sqrt{4940 - 3600}}$$

$\Sigma XY = 288$

$$r = \frac{480}{\sqrt{520 \times 1340}}$$

$\Sigma X^2 = 212$

$$r = \frac{480}{\sqrt{696800}} = \frac{480}{834.74} = 0.58$$

$\Sigma Y^2 = 494$

$r = 0.58$

The co-efficient correlation between X and Y is 0.58.

Positive and moderately related to each other.

Karl Pearson's Product Movement Correlation

$$\gamma = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

Find Product Moment Correlation for the Data

X	**Y**
5	8
7	9
3	5
1	4
9	9
12	13
8	7
3	9
ΣX =48	ΣY =64

Calculation of Product Moment Correlation Co-Efficient

X	Y	X^2	Y^2	XY
5	8	25	64	40
7	9	49	81	63
3	5	9	25	15
1	4	1	60	4
9	9	81	81	81
12	13	144	169	156
8	7	64	49	56
3	9	9	81	27
ΣX =48	ΣY =64	ΣX^2 =382	ΣY^2 =566	ΣXY =442

$$\gamma = \frac{8\times 442-(48)\times(64)}{\sqrt{8\times 382-(48)^2}\sqrt{8\times 566-(64)^2}}$$

$$\frac{3536-3072}{\sqrt{3056-2304}\sqrt{4528-4096}} = \frac{464}{\sqrt{752}\sqrt{432}}$$

$$= \frac{464}{27.42\times 20.78} = \frac{464}{569.8} = 0.814$$

γ =0.814

Very high positive correlation.

Find Product Moment Correlation for the Data

X	Y
52	38
57	36
50	30
54	52
59	37
60	35

Calculation of Product Moment Correlation Co-Efficient

X	Y	X (X – 50)	Y (Y-30)	X^2	Y^2	XY
52	38	2	8	4	64	16
57	36	7	6	49	36	42
50	30	0	0	0	0	0
54	52	4	22	16	484	88
59	37	9	7	81	49	63
60	35	10	5	100	25	50
		ΣX =32	ΣY =48	ΣX^2 =250	ΣY^2 =658	ΣXY =259

$$\gamma = \frac{6\times 259 - 32\times 48}{\sqrt{6\times 250-(32)^2}\sqrt{6\times 658-(48)^2}}$$

$$= \frac{1554-1536}{\sqrt{1500-1024}\sqrt{3948-2304}}$$

$$= \frac{18}{\sqrt{476}\sqrt{1644}}$$

$$= \frac{18}{21.82\times 40.55} = \frac{18}{884.80} = 0.020$$

γ =0.020

Very low correlation.

Find Product Moment Correlation for the Data

X	Y
6	5
3	2
4	3
5	7
2	4
8	9
4	3

3	5
5	6
7	8
ΣX =47	ΣY =52

Calculation of Product Moment Correlation Co-Efficient

X	**Y**	**X^2**	**Y^2**	**XY**
6	5	36	25	30
3	2	9	4	6
4	3	16	9	12
5	7	25	49	35
2	4	4	16	8
8	9	64	81	72
4	3	16	9	12
3	5	9	25	15
5	6	25	36	30
7	8	49	64	56
ΣX =47	ΣY =52	ΣX^2 =253	ΣY^2 =318	ΣXY =276

$$\gamma = \frac{10\times 276 - 47\times 52}{\sqrt{10\times 253-(47)^2}\sqrt{10\times 318-(52)^2}}$$

$$= \frac{2760-2444}{\sqrt{2530-2209}\sqrt{3180-2704}}$$

$$= \frac{316}{\sqrt{321}\sqrt{476}} = \frac{316}{17.92\times 21.82}$$

$$= \frac{316}{391.01} = 0.808 = 0.81$$

γ = 0.81

Very high correlation.

Uses advantages and disadvantages of Pearson's correlation

The Pearson's product moment correlation is most popular and widely used in research and measurement.

The following are the main uses and advantage

1. The reliability of test is calculated in terms of Pearson (r)
2. The validity is estimated by the co-efficient of correlation (r)
3. The cut of score is determined empirically with the help of scatter plot.
4. Item discrimination power is calculated by using Pearsons (r)
5. Multiple correlation based on Pearson's r
6. Partial correlation employs the co-efficient of correlation (r)
7. Factor-analysis technique is the extension of Pearson's r
8. In guidance service specially for prediction purpose Pearson's r is used.
9. The theories of intelligence have been developed by using the co-efficient of correlation of Pearson's r
10. Most of the personality theories are also developed by using this correlation.
11. The correlational studies in behavioural sciences commonly employ the Pearson's product moment correlation technique.

Disadvantages

1. It is a linear correlation. When the two variables have the linear distribution would yield accurate co-efficient of correlation, but the two variables are curvelinearly distributed. Therefore the correlation of co-efficient of two variables is not dependable. This assumption is taken into consideration while using this technique.
2. The distribution of scores of the two variables should be normal. It the distributions are skewed, it would not yield dependable correlation. The assumption is not usually observed.

BISERIAL CORRELATION

To estimate the relationship between a continuous variable and a dichotomous variable. The term dischotomous means cut into two pads. The variable of social adjustment can be dischotomised as "socially adjusted subjects" and socially maladjusted subjects.

The formula for biserial correlation is

$$rbis \frac{M_p - M_q}{\sigma_t} \times \frac{pq}{y}$$

r bis-biserial r

Mp & Mq-mean test scores respectively for those who pass and fail the item

P&Q - Proportions who pass and fail the item

y-height of the ordinate of the normal curve at the point of division between p and q proportions

σ_t -S.D

A rehearsal group and a non-rehearsal group obtained the following scores on their performance. Find out the correlation between rehearsal and performance is distribution.

Scores	**Rehearsal**	**Non-rehearsal**
90-99	3	9
80-89	6	15
70-79	5	18
60-69	12	36
50-59	10	18
40-49	8	30
30-39	6	24
Total	50	200

Mp-60.8

Mq-59.5

S.D-17.63

P=50/200 = 0.25

Q=1-p

=1-0.25=0.75

$$rois = \frac{60.9 - 59.5}{17.63} \times \frac{0.25 \times 0.75}{0.318}$$

=.047

The correlation is negligible.

POINT BISERIAL CORRELATION

When test items are scored simply as 1 if correct, and o if incorrect, that is right or wrong, the assumption of normality in the distribution of right wrong responses is generally not met. Other example of genuine or natural dichotomies are male-females, rural-urban, living-dead, loyal-disloyal. In such cases, the point biserial correlation (rpbis) instead of rbis can be used.

Calculate point biserial correlation for the following data.

Test criterion	Item No. (y)
15	1
14	0
13	0
15	1
10	1
15	1
13	0
12	1
15	1
10	0
11	0
Σ 143	6

N_1-No. of passing = 6

N_2-No of failures = 5

M_1-mean for the passing item

M_2-mean for the failing item

$$M_1 = \frac{82}{6} = 13.66$$

$$M_2 = \frac{61}{5} = 12.2$$

$$M_{tot} = \frac{143}{11} = 13$$

$$S.D = \frac{1}{N}\sqrt{N\Sigma X^2 - (\Sigma X)^2}$$

$$= \frac{1}{11}\sqrt{11(1899) - (143)^2}$$

$$= \frac{1}{11}\sqrt{20{,}899 - 20{,}449}$$

$$= \frac{1}{11}\sqrt{440}$$

$= \frac{1}{11} \times 20.98$

=1.91

$P = \frac{6}{11} \times 0.54$

$Q = \frac{5}{11} \times 0.45$

$rpbi = \frac{M_p - M_q}{\sigma_t} \sqrt{pq}$

$= \frac{13.67 - 12.20}{1.91} \sqrt{0.55 \times 0.45}$

$= \frac{1.47}{1.91} \sqrt{0.2475}$

$= \frac{1.47}{1.91} \times 0.497$

rpbi=0.38

TETRA CHORIC CORRELATION

The biserial and point biserial correlations were used when one variable was continous and expressed as test scores and the other variable was dichotomous. When both the variables are dichotomous, we cannot calculate point biserial or biserial correlation. For such variations, trtrachoric correlation rt can be calculated.

$$r_t = \cos\left(\frac{180^0 \times \sqrt{BC}}{\sqrt{AD} + \sqrt{BC}} \right)$$

Find the tetrachotic correlation for the following data.

	Fail	Pass	
Trained	40B	20A	60
Untrained	22D	15C	37
Total	62	35	97

$$r_t = \cos\left(\frac{180^0 \times \sqrt{BC}}{\sqrt{AD} + \sqrt{BC}} \right)$$

$$= \cos \frac{180^0 \times \sqrt{600}}{\sqrt{440} + \sqrt{600}}$$

$$= \cos\left(\frac{3775.71}{45.47}\right)$$

$= \text{Cos } 83^0$

$r_t = 0.122$

4.18 PHI CO- EFFICIENT (OR) ϕ COEFFICIENT

This technique of correlation ϕ is used when both the variables have the true dichotomy, such as true-false, right-wrong. ie, two categories of each variable.

$$\phi = \frac{AD-BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

ITEM X

	Wrong	**Right**	**Total**
Right	20 B	80 A	A+B 100
Wrong	70 D	55 C	C+D 125
Total	B+D	A+C	225 N

$$\phi = \frac{AD-BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

$$= \frac{80 \times 70 - 55 \times 20}{\sqrt{100 \times 125 \times 90 \times 135}}$$

$$= \frac{5600 - 1100}{\sqrt{151,87500}}$$

$$= \frac{4,500}{12,323.757}$$

= 0.36

PARTIAL CORRELATION

In many research studies data as more than two variables are collected and forms of multivariate analysis are required. For determining the relationship between two of the variables, when the influence of third variable is eliminated or partialed out, this is the net correlation or multiple regression.

It is well known fact that if two variables are correlates with a third, the relationship between the variables is influenced by the third variable. For example, we measure the academic achievement and the intelligence of a group of normal children in a normal situation. We may find that the academic achievement and intelligence are highly correlated. The correlation between two variables academic achievement and intelligence, may be due to third variable is age. This is true that as age increases of their children, their academic achievement and intelligence will increase. If the influence of age is removed or eliminated not correlation between academic achievement and intelligence may be obtained. Partial correlation eliminates the effect of third variable to obtain the net correlation. The partial correlation technique relates in determining the correlation between two of the variables when the influence of other is removed. It is the net correlation or partial correlation let us take X_1 as academic achievement, X_2 as intelligence and X_3 as age of the children. The correlation between X_1 and X_2 is denoted by r_{12} and the effect of X_3 is eliminates by statistical method is known as partial correlation co-efficient and is written as $r_{12.3}$. The formula for calculating partial correlation is

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)}\sqrt{(1-r_{23}^2)}}$$

Where r_{12} is Pearson's correlation between X_1 and X_2.

r_{13} is the Pearson's correlation between X_1 and X_3.

r_{23} is the Pearson's correlation between X_2 and X_3.

Where $r_{12.3}$ is the co-efficient of partial correlation between X_1 and X_2 keeping X_3 constant.

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1-r_{12}^2)}\sqrt{(1-r_{23}^2)}}$$

where $r_{13.2}$ is the co-efficient of partial correlation between X_1 and X_3 keeping X_2 constant.

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1-r_{12}^2)}\sqrt{(1-r_{13}^2)}}$$

where $r_{23.1}$ is the co-efficient of partial correlation between X_2 and X_3 keeping X_1 constant.

Thus, for three variables X_1, X_2 and X_3 there will be three co-efficient of partial correlation each studying the relationship between two variables when the third is held constant. Thus in determining a partial correlation co-efficient between X_1 and X_2, we attempt to remove the influence of X3 from each of the two variables.

Zero order, first order and second order co-efficients

Partial co-efficients such as $r_{12.3}$, $r_{13.2}$, $r_{23.1}$ are referred as first order co-efficients. This is because are variable has been held constant. Simple correlation co-efficients correlation between two variables are called zero order co-efficients, since no variables are held constant.

$r_{12.34}$, $r_{13.24}$, $r_{23.41}$ etc, are called second order co-efficients since two variables are kept constant.

Generally it is stated that the order of co-efficient indicates the number of variables that have been held constant statistically.

The following zero-order correlation co-efficients are given: $r_{12} = 0.98$, $r_{13} = 0.44$ and $r_{23} = 0.54$ calculate the partial correlation between first and the third variables keeping the effect of second variable constant.

Solution

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1-r_{13}^2)}\sqrt{(1-r_{23}^2)}}$$

$$r_{13.2} = \frac{0.44 - (0.98)(0.54)}{\sqrt{1-(0.98)^2}\sqrt{1-(0.54)^2}}$$

$$r_{13.2} = \frac{0.44 - 0.5292}{\sqrt{1-0.9607}\sqrt{1-0.2916}}$$

$$r_{13.2} = \frac{-0.0892}{\sqrt{0.0396}\sqrt{0.7084}}$$

$$= \frac{-0.0892}{0.199 \times 0.842}$$

$$= \frac{-0.0892}{0.1676}$$

$$r_{13.2} = -0.532$$

There is a negative correlation between the variables 1 and 3 and keeping second variable as constant.

Example: On the basis of the test results, the correlation between academic achievement (X_1) and Intelligence (X_2), correlation between academic achievement (X_1) and self-concept (X_3) and correlation between Intelligence (X_2) and self-concept (X_3) of the students are found to be.

$r_{12} = 0.8$ $r_{13} = 0.65$ and $r_{23} = 0.7$

compute the partial correlation between academic achievement and intelligence of the students keeping self concept as constant.

Solution

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)}\sqrt{(1-r_{23}^2)}}$$

$$r_{12.3} = \frac{0.8 - (0.65)(0.7)}{\sqrt{1-(0.65)^2}\sqrt{1-(0.7)^2}}$$

$$r_{12.3} = \frac{0.8 - 0.455}{\sqrt{1-0.4225}\sqrt{1-0.49}}$$

$$= \frac{0.345}{0.76 \times 0.714}$$

$$= \frac{0.345}{0.543}$$

$$r_{12.3} = 0.635$$

Partial Correlation Co-Efficients in case of Four Variables

When four variables are involved in a correlation problem, there are twelve possible first order co-efficients. Some of them as:

$$r_{14.2} = \frac{r_{14} - r_{12}r_{24}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{24}^2}}$$

$$r_{14.3} = \frac{r_{14} - r_{13}r_{34}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{34}^2}}$$

Second Order Partial Correlation Co-Efficients

Second order co-efficients may be obtained from first order co-efficients. In case of four variables. If $r_{12.34}$ is the co-efficients of partial correlation between X_1 and X_2 keeping X_3 and X_4 constant, then

$$r_{12.34} = \frac{r_{12.4} - r_{13.4}r_{23.4}}{\sqrt{1-r_{13.4}^2}\sqrt{1-r_{23.4}^2}}$$

Similarly

$$r_{13.24} = \frac{r_{13.4} - r_{12.4}r_{23.4}}{\sqrt{1-r_{12.4}^2}\sqrt{1-r_{23.4}^2}}$$

and

$$r_{14.23} = \frac{r_{14.3} - r_{12.3}r_{24.3}}{\sqrt{1-r_{1.23}^2}\sqrt{1-r_{24.3}^2}}$$

The value of a partial correlation co-efficient is usually interpreted (via) the corresponding co-efficient of partial determination, which is merely the square of the formes. Thus if $r_{12.3} = 0.4$ then

$r_{12.3}^2 = 0.16$.

The function of partial correlation analysis is the measurement of relationship between two facts with the effects of one or more other factors eliminated.

ADVANTAGES OF CORRELATION

1. It is useful in correlational studies
2. It is used in finding validity of the test
3. It is used in reliability of a test
4. It is useful to write down the regression equation

REGRESSION

It was introduced by Sir Francis Galton in 1877 in his study of heredity.

The term regression has been from the word 'to regress' which means tendency to go back. This statistical method is employed for predicting or estimating the unknown value called dependent variable, from value of another variable is known as independent variable.

Definition and Importance

1. According to Balir, "Regression is the measure of the average relationship between two or more variables in terms of the original units of the data".
2. According to Wallis and Robert, "regression is often more important to find out what is the actual relation is, in order to estimate or predict one variable and statistical technique appropriate in such a cases is called regression".
3. According to Ya-Lu-Chow, "Regression analysis attempt to establish the nature of the relationship between the variables that is to study the functional relationship between the variables and thereby provide a mechanism for prediction or forecasting".
4. In regression analysis; the independent variable is known as the "regression" or "predictor" or "explanatory" and the dependent variable is known as "regressed" or "explained" variable. With the help of regression analysis, we can predict the unknown value of one variable from the known value of another variable.

4.21.1. Regression Equations

a. Regression equation of X on Y is in the form: $X = r (Sx / Sy) (Y- My) + Mx$

b. Regression equation of Y on X is in the form: $Y = r (Sy / Sx) (X-Mx) + My$

Where as;

X and Y are variables.

Mx and My are Mean for X and Y.

Sx and Sy are Standard Deviation for X and Y.

r is the correlation coefficient of X and Y.

Slope is $Y = Mx + C$

Two achievement tests in maths and science are administered for the sample and following statistics were obtained.

	Maths (X)	Science (Y)
Mean	60	48
S.D	15	12
Rxy	0.60	

The science score of a student is 40. Predict his / her maths score and the maths score of another student is 70. Predict his / her science score.

Regression Equation of X on Y is $X = r(Sx / Sy) (Y-My)+Mx$

$= 0.6 \times (15/12) \times (40-48) + 60$

$= (90/12) \times (-8) + 60$

$= (0.75) \times (-8) + 60$

$= 54$

Maths score of the student is 54.

Regression equation of Y on X is $Y = r(Sy / Sx) (X-Mx) + My$

$= 0.6 \times (12/15) \times (70-60) + 48$

$= (7.2/15) \times (10) + 48$

$= (0.48) \times (10) + 48$

$= 52.8$

Science score of the student is 52.8

4.21.2. Advantages and Disadvantages of Regression

1. It is easy to understand and intercept.
2. It is very for prediction purposes.
3. The regression line is easily and quickly obtained.
4. For purely descriptive purposes, it is sufficiently accurate.
5. It is used more than the correlation analysis in many scientific studies.
6. It predicts the values of dependent variables from the values of the independent variable.
7. We can calculate coefficient correlation (r) and coefficient of determination (r2) with the help of regression coefficient.

Disadvantages

1. It is not represented by any algebraic equation.
2. It is subjective; therefore different persons would draw different lines from the same set of data.

Educational Uses

1. It is used in diagnostic test.
2. It is used in vocational guidance.
3. It is used in educational guidance.
4. It is used for selecting the right person for the job.
5. It is used for analyzing predictive validity of the tests.

4.19.4.DIFFERENCE BETWEEN CORRELATION AND REGRESSION

Correlation	Regression
1. Correlation is the relationship between two or more variables, which vary in sympathy with the other in the same or the opposite direction.	Regression means going back and it is a mathematical measure showing the average relationship between two variables.
2. Both the variables X and Y are random variables.	Here X is a random variable and Y is the fixed variable. Some times both the variables may be random variables.
3. It finds out the degrees of relationship between two variables and not the cause and effect of the variables.	It indicates the cause and effect relationship between the variables and establishes a functional relationship.
4. It is used for testing and verifying the relation between variables and gives limited information.	Besides verification, it is used for the prediction of on value, in relationship to the other given value.
5. The coefficient of correlation is a relative measure. The range of relationship lies between + / -1.	Regression coefficient is an absolute figure. If we know the value of the independent variable, we can find the value of the dependent variable.
6. There may be nonsense correlation between two variables.	In regression there is no such nonsense regression.
7. It has limited application because it is confined only to linear relationship between the variables.	It has wider application, as it studies linear and non-linear relationship between the variables.
8. It is not very useful for further mathematical treatment.	It is widely used for further mathematical treatment.
9. If the coefficient of correlation is positive, then the two variables are positively correlated and vice versa.	The regression coefficient explains that the decrease in one variable is associated with the increase in the other variable.

DESCRIPTION OF THE NORMAL PROBABILITY CURVE

Normal distribution is the cornerstone of modern statistics, which is highly useful in statistic and is need to deal with "continuous probability distribution". The normal distribution was first discovered by De-moivre in 1733 to solve problems in games or chances. Later it was applied in natural and social science by the French Mathematician La Place (1949).

The limiting frequency curve obtained as 'n' becomes large is called 'normal curve ' or normal probability curve' or' normal frequency curve', i.e., the graph of the normal distribution is called normal curve, which is a bell-shaped curve expanding in both the

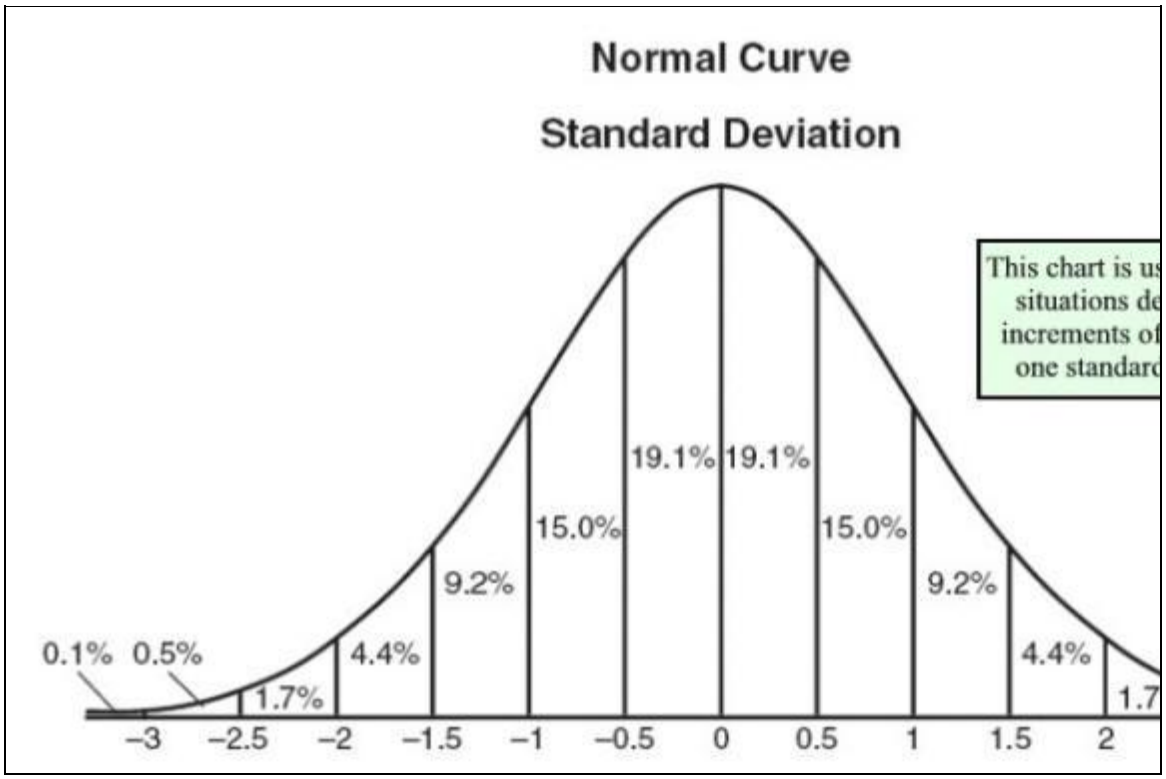
directions, arriving nearer and nearer to the horizontal axis but never touches it. Normal distribution is not an actual distribution of scores on any test of ability or academic achievement, but is instead a mathematical model. It is of great value in educational research when we make use of mental measurement.

The normal distribution is determined by the parameter-mean and standard deviation. For different values of mean and standard deviations, we get different normal distributions. The area under the normal curve is always taken as unity so as the represent total probability. The area is of great importance in a variety of problems because such an area represents frequency.

J.P. Guilford has defined normal probability curve comprehensively.

"Normal probability curve is well defined, well structure, mathematical curve, having a distribution of the scores with mean, median and mode are equal. It does not occur in nature. It is not a biological or psychological curve".

Normal Probability Curve



5.3.1. Properties of Normal Probability Curve

1. The normal curve is symmetrical about the mean. The distribution of the frequency on either side of the maximum ordinate of the curve is exactly the same. It is a bell shaped curve.
2. The value of mean, median and mode will coincide because the distribution is symmetrical.
3. The height of the normal curve is maximum at the mean, and in the unit normal curve is equal to 0.3989.
4. Mean and standard deviations are the parameters of normal distribution.
5. It has only one mode, since, there is only one maximum point.
6. The curve is asymptotic. It approaches but does not meet horizontal axis (base line) and extends from - ∞ (minus infinity) to + ∞ (plus infinity).
7. Since the distribution is symmetrical the coefficient of skewness (β_1) = 0; and kurtosis (β_2) = 3.

8. The mean deviation is equal to 0.7979 σ = 4/5 σ. The quartile deviation is 0.6745 σ.

9. The points of inflection occur at $\bar{x} \pm \sigma$.

10. The limits of the normal curve are - ∞ and + ∞.

Area between – 1 σ to + 1 σ is 68.26%.

The total area of a normal curve is 1.

A normal curve with mean "0" and standard deviation (σ) = 1 is known as the standard normal curve.

Uses of the Normal Probability curve

There are number of applications of the normal curve in the field of educational research.

1. To convert raw scores into standard scores.

Suppose the student A gets 65 marks in a test in science administered to 100 students. The mean and standard deviation of the test marks for the group are 50 and 5 respectively. Thus the standard score of student A is

$$Z = \frac{X - M}{S} = \frac{65 - 50}{5} = \frac{15}{5} = 3$$

This indicates the score of 65 is 3 standard deviation above the mean

2. To calculate the percentile rank of scores

When scores in a distribution are assumed to be normally distributed, percentile rank can be calculated from the Z score.

3. To scale responses to opinionnaires, judgement, ratings or rankings by transforming them numerical values.

4. To normalize a frequency distribution. It is an important step in standardizing a psychological test or inventory.

5. To test the significance of observations in experiments, findings them relationships with the chance fluctuations or errors that are result of sampling procedures.
6. To generalize about population from which the samples are drawn by calculating the standard error of mean and other statistics.
7. To compare two distributions. The NPC is used to compare two distributions.
8. To determine the difficulty values. The Z scores are used to determine the difficulty values of test items.
9. To classify the groups.

The Normal Probability Curve (NPC) is used for classifying the groups and assigning grades to individuals.

10. To determine the level of significance. The levels of significance of statistics results are determined in terms of NPC limits.

Applications of Normal Probability Curve

Find the area under the normal curve for Z = 1.67

Solution

From the table, we find the area for Z = 1.67 as .4525.

Find the area to the left of Z = 2.17

Area to the left of 0 is 0.5

Area between 0 and 2.176, is .4850

Therefore the area left of Z = 2.176 is

0.4850+.5000 = .9850

.

Find the area between Z = 0 and Z = -1.54. This has been shown in the following figure.

A normal curve has mean 20 and standard deviation 10. Find the area between $X_1 = 15$ and $X_2 = 40$

Solution

$$Z_1 = \frac{X_1 - M}{S} = \frac{15-20}{10} = \frac{-5}{10} = -0.5$$

$$Z_2 = \frac{X_2 - M}{S} = \frac{40-20}{10} = \frac{20}{10} = 2$$

The area corresponding to is

Z_1 = -0.5 is 0.1915 and

Z_2 = 2 is 0.4772

$\therefore$ Area between X_1 = 15 and X_2 = 20 is

0.195 + 0.4772 = 0.6687

In an intelligence test administered to 1000 students the score was 42 and standard deviation was 24.

Find i) number of students exceeding the score 50, ii) number of students getting scores between 30 and 54.

Solution

Given M = 42, S = 24

X = 50

$$Z = \frac{X - M}{S} = \frac{50 - 42}{24} = \frac{8}{24} = 0.33$$

Area between o and 33 is 0.1293

Area to the right of Z = 0.33 is

= 0.5-0.1293

=0.3707

This has been shown in the following figure.

Number of students exceeding the score 50 = .3707 x 1000 = 3707 or 371

ii) Number of students lying between 30 and 54.

$X_1 = 30$ $X_2 = 54$ M = 42 S = 24

$$Z_1 = \frac{X_1 - M}{S} = \frac{30 - 42}{24} = -\frac{12}{24} = -0.5$$

$$Z_2 = \frac{X_2 - M}{S} = \frac{54 - 42}{24} = \frac{12}{24} = 0.5$$

Area between Z = 0 and Z_1 = -0.5 is 0.1915

Area between Z = 0 and Z_2 = 0.5 is 0.1915

Area between Z_1 = -0.5 and Z_2 = 0.5 is 0.3830

This has been shown in the figure

The number of students lying between 30 and 54 =0.3830 x 1000 = 383

In a normal distribution 31% of the items are under 45 and 8% are over 64. Find the mean and standard deviation of the distribution.

Solution

31% are under 45. (Area to the left is 0.31, but right from this point, 0.31 to 0.5 is 0.19, the $Z_1 = \frac{45 - M}{S} = -0.5$ value corresponding to Z = 0.5.

45-M = -0.53

8% of the items are above 64. The area to the rights of the ordinate at M = 64 to the mean ordinate is 0.5-0.08 = 0.42. The value corresponding to this 1.4 (from the tables)

$Z_2 = \frac{64 - M}{S} = 1.4$

=64-M = 1.4 S -2

Subtract 2 from 1

-19 = -1.901 S

S = $\frac{-19}{-1.90}$ =9.99 or 10.

Substitute the value in 2

64 – M = 1.405 x 10

= 14.05

-M = -49.95

M = 49.95

= 50.

5.3.3 Skewness

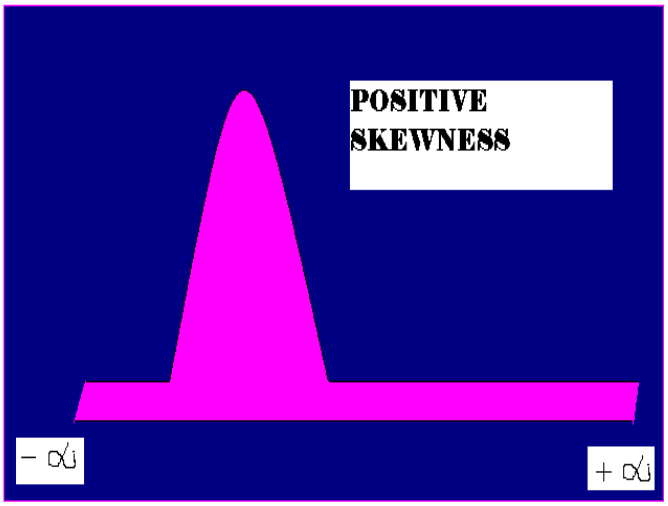
In everyday language, the terms “skewed” and “askew” are used to refer to something that is out of line or distorted on one side. When referring to the shape of frequency or probability distributions, “skewness” refers to asymmetry of the distribution. A distribution with an asymmetric tail extending out to the right is referred to as “positively skewed” or “skewed to the right”, while a distribution with an asymmetric tail extending out to the left is referred to as “negatively skewed” or “skewed to the left”. Skewness can range from minus infinity to positive infinity.

That is,Skewness is refers to lack of symmetry in a normal curve. There are two types of skwness they are,

1. Positive Skewness and
2. Negative skewness.

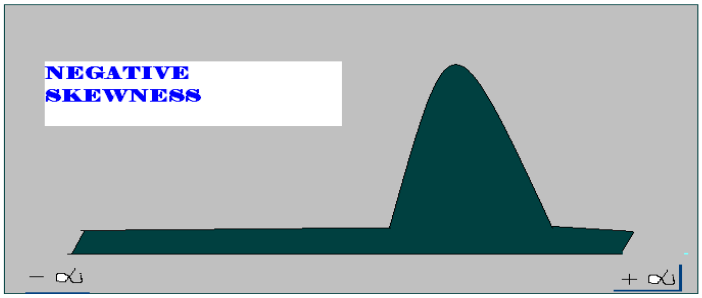
Positive skewness

- When the scores are massed at low end of the scale and a long tail is running along the high end of the scale, the distribution is said to be positively skewed.



Negative Skewness

When the scores are massed at the high end of the scale and a long tail is running along the low end of the scale, the distribution is said to be negatively skewed.

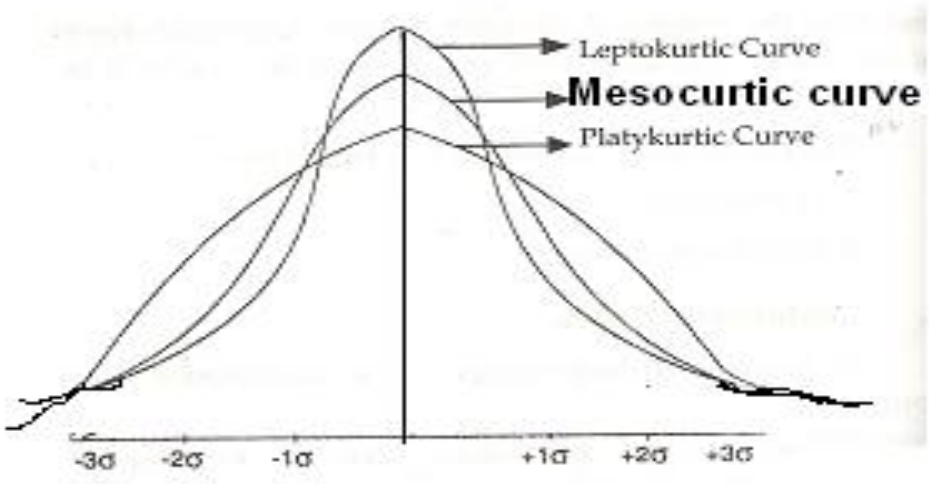


Skewness is calculated by using the following formula

$$Skewness = \frac{3(Mean - Median)}{SD}$$

5.3.4. Kurtosis

Kurtosis refers to peakness or flatness of a normal curve. A normal curve is called as Mesokurtic. A curve which is more peak than the normal curve is called Leptokurtic. A curve, which is flatter than the normal curve, which is called Platykurtic.



STANDARD ERROR

Statistical inference treats two different classes of problems. They are hypothesis testing and estimation. Hypothesis testing is to test some hypothesis about present population from which the sample is drawn. Estimation is to use the 'statistics' obtained from the sample as estimate of the unknown parameter of the population from which the sample is drawn.

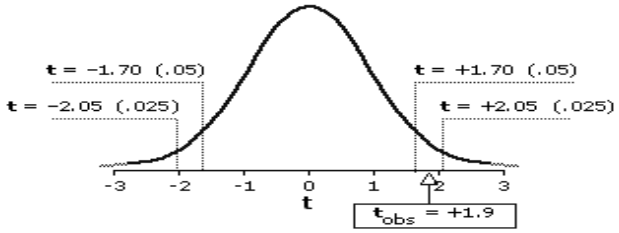
The standard deviation of the sampling distribution is called the standard error.

The word 'error' is used in place of 'deviation' to emphasize that variation among sample means is due to sampling errors.

Standard error is so called because it measures the sampling variability due to chance or random forces. Hence to clarify the term standard error it is necessary to describe sampling distribution. If we select a number of independent random samples of a definite size from a given population and calculated some statistics (like the mean, standard deviation etc) from each sample, we shall get a series of values of these statistics or functions. These values obtained from the different samples can be put in the form of a frequency distribution. The distribution so formed of all possible values of a statistics is called the sampling distribution or the probability distribution of that statistic. Thus if we draw, 100 random samples from a given population and calculate their means. We shall

get a series of 100 means which would form a frequency distribution. This distribution will be known as the sampling distribution of the means.

Sampling Distribution



	Level of Significance for a Directional Test				
	.05	.025	.01	.005	.0005
	Level of Significance for a Non-Directional Test				
	---	.05	.02	.01	.001
df = 28	1.70	2.05	2.47	2.76	3.67

5.4.1. Utility of Concept of Standard Error

Standard error plays a very important role in the large sample theory and forms the basic of the testing of hypothesis.

1. It is used as an instrument in testing a given hypothesis.
2. Standard error provides an idea about the unreliability of a sample. The greater the standard error, the greater is the departure of actual frequencies from the expected areas and hence the greater the unreliability of the sample. The reciprocal of standard error i.e., is a measure of reliability (or precision) of the sample.
3. Standard error (S.E) can be used to determine the confidence limits for population values like mean, population and standard deviation, (with the help of S.E one can determine the limits within which the parameter values are expected to lie. This is made possible because for larger samples, sampling distribution tend to approximate a normal distribution.

5.4.2. Standard Error of the mean

Standard error of the mean measured only sampling errors. Sampling errors are errors involved in estimating a population parameters from a sample instead of including all the essential information in the population.

1. When standard deviation of the population is known

$$S.E\,\bar{x} = \frac{\sigma p}{\sqrt{n}}$$

Where S.E $\bar{x}$ refers to the standard error of the mean

σp = Standard deviation of the population

n = Number of observations in the sample

σ = Standard deviation of the sample

2. When the standard deviation of the population is not known, we have to use standard deviation of the sample in calculating standard error of mean. The formula for calculating standard error is

$$S.E\,\bar{x} = \frac{\sigma}{\sqrt{n}}$$

Illustration

The mean height obtained from a random sample of size 100 is 64 inches. The standard deviation of the distribution of height of the population is known to be 3 inches. Test the statement that the mean height of the population is 67 inches at 5% level of significance. Also set up 99% limits of the mean height of the population.

Solution: Let us take the hypothesis that there is no significant difference between the sample mean and the population mean.

$$S.E\,\bar{x} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{100}} = 0.3$$

$$\frac{Diff}{S.E} = \frac{67-64}{0.3} = 10$$

Since the difference is more than 1.96 S.E (5% level) the hypothesis is rejected. Hence the mean height of the population could not be 67 inches 99% probable limits of the mean height of the population.

$$= \bar{x} \pm 2.58\,S.E$$
$$= 64 \pm 2.58(.3)$$
$$= 64 \pm 0.774$$
$$= 63.2\,to\,64.8$$

Two-tailed test for Difference between the means of two samples

1. If two independent random sample with n1 and n2 numbers (both sample sizes one greater than 30) respectively are drawn from the same population of standard deviation σ^1 the standard error of the difference between the sample means is given by the formula.

$$t = \frac{M_1 - M_2}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}}$$

Where

M_1 = Mean for the I group

M_2 = Mean for the II group

S_1 = S. D for the I group

S_2 = S. D. for the II group

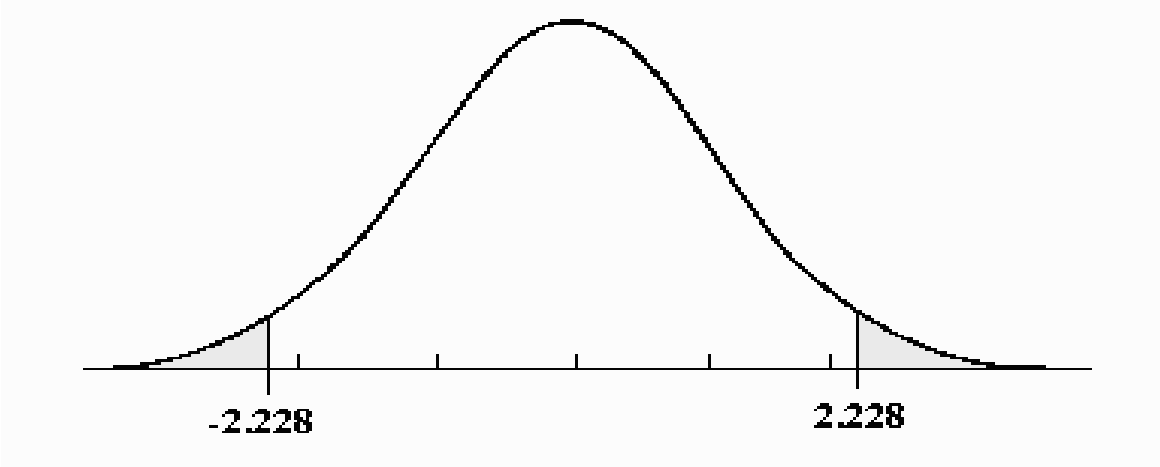
N_1 = Size of the I group

N_2 = Size of the II group

5.4.3. Two-Tailed T-Tests

A two-tailed t-test divides a curve in half, placing half in the each tail. The null hypothesis in this case is a particular value, and there are two alternative hypotheses, one positive and one negative. The critical value of t, tcrit, is written with both a plus and minus sign (±). For example, the critical value of t when there are ten degrees of freedom (df=10) and a is set to .05, is tcrit= ± 2.228. The sampling distribution model used in a two-tailed t-test is illustrated below:

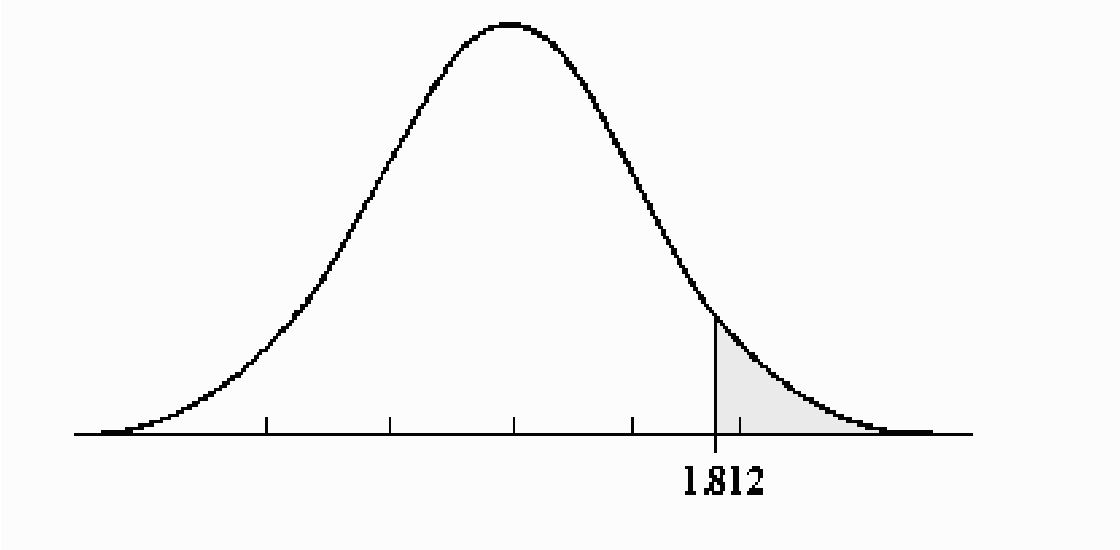
TWO-TAILED t-TEST



- There are really two different one-tailed t-tests, one for each tail. In a one-tailed t-test, all the area associated with a curve is placed in either one tail or the other. Selection of the tail depends upon which direction jobs would be (+ or -) if the

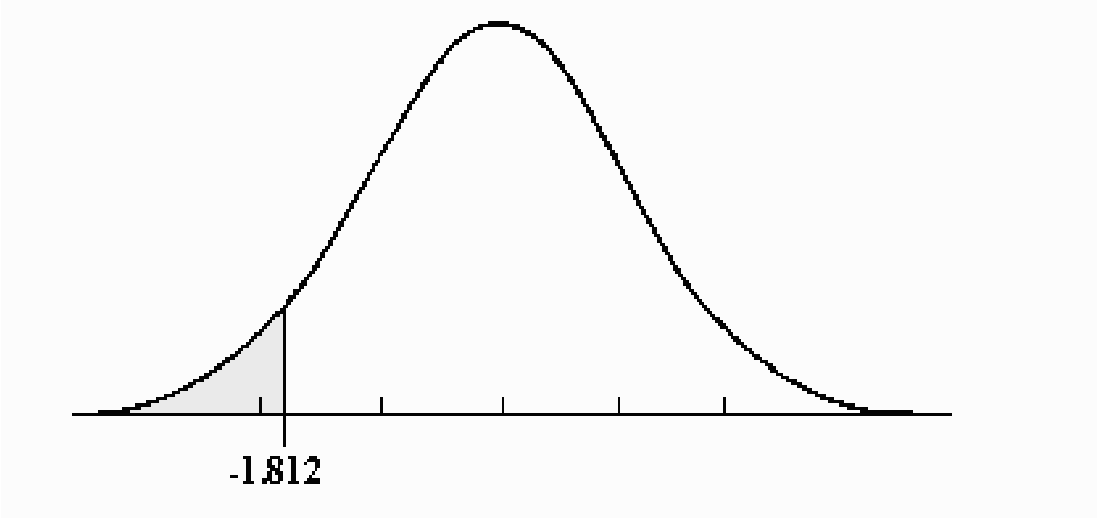
results of the experiment came out as expected. The selection of the tail must be made before the experiment is conducted and analyzed.

- **A one-tailed t-test in the positive direction is illustrated below**



A one-tailed t-test

A one-tailed t-test in the negative direction is illustrated below



5.4.4. Comparison of One and Two-tailed t-tests

- 1. If tOBS = 3.37, then significance would be found in the two-tailed and the *positive* one-tailed t-tests. The one-tailed t-test in the negative direction would not be significant, because was placed in the wrong tail. This is the danger of a one-tailed t-test.
- 2. If tOBS = -1.92, then significance would only be found in the *negative* one-tailed t-test. If the correct direction is selected, it can be seen that one is more likely to reject the null hypothesis. The significance test is said to have greater *power* in this case.
- The selection of a one or two-tailed t-test must be made before the experiment is performed.

Sigma Value for rejection of hypothesis

Sigma Value for rejection of hypothesis

Statistical Tests	Level of signification at 0.05	Level of Significance at 0.01
One tailed Test	1.64	2.33
Two tailed Test	1.96	2.58
Probability	0.95	0.99

Student's Distribution 't'

- To compute t-value for the significance of the difference between two means, when N is fewer than 30, the formula is:

$$t = \frac{(M_1 - M_2)}{\sqrt{\frac{(N_1 - 1)S_1^2 + (N_2 - 1)S_2^2}{(N_1 + N_2 - 2)}}} \sqrt{\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$$

For the test of significance of the difference between two means the number of degrees of freedom would be:

(N1-1) + (N2-1) = (N1+N2-2) =df

Two Types of Error (α and β errors)

- The two types of error may arise in taking a decision about the null hypothesis-
- 1. When the null hypothesis is true, an alternative hypothesis may be accepted. It is known as α or Type 1 error.
- 2. When an alternative hypothesis is true, the null hypothesis may be accepted. This is called β error or Type II error

5.4.5. Significance of the Difference Between Two Means

- There are two situations for ascertaining the significance of the difference between two means.
- 1. Means of two independent groups, uncorrelated means and
- 2. Means of the same group in different two conditions-correlated means.
- The 't' test is applied in both the situation but it employs different methods for ascertaining the significance of different of two means

Small Samples, significance of difference of two means

$$t = \frac{M_1 - M_2}{S\sqrt{\left(\frac{N_1 + N_2}{N_1 N_2}\right)}} \; with \; df = (N_1 + N_2 - 2)$$

Large Samples-Significance of Mean Difference

S.E of the difference between sample means

$$= \sqrt{\sigma^2 \left[\frac{1}{n_1} + \frac{1}{n_2} \right]}$$

If σ is unknown, sample standard deviation for combined samples must be substituted.

2. If two random samples with $\bar{x}_1$, σ^1, n_1 and $\bar{x}_2$, σ^2, n_2 respectively are drawn from different population, then the S.E of the difference between the mean is given by the formula.

$$= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

and where σ^1 and σ^2 are unknown.

S.E of the difference between means

$$= \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Where S_1 and S_2 represent standard deviations of the two samples.

The null hypothesis to be tended is that there is no significant difference in the means of the two samples. i.e.,

$H_0 : \mu_1 = \mu_2$ ⟵ null hypothesis, there is no difference.

$H_1 : \mu_1 \neq \mu_2$ ⟵ alternative hypothesis, a difference exists.

Illustration

A study was conducted on social anxiety. The data were collected from male and female students.

Group	Mean	SD	N
Male students	69.32	11.56	144
Female students	65.95	9.88	256

Verify if there is any significant difference between male and female students in their social anxiety.

Solution

Null hypothesis: There is no significant difference between male students and female students in their social anxiety.

$$t = \frac{M_1 - M_2}{\sqrt{\frac{S_1^{\,2}}{N_1} + \frac{S_2^{\,2}}{N_2}}}$$

$$= \frac{69.32 - 62.95}{\sqrt{\frac{(11.56)^2}{144} + \frac{(9.88)^2}{256}}}$$

$$= \frac{3.37}{\sqrt{0.93 + 0.38}}$$

$$= \frac{3.37}{1.31}$$

= 2.956

At 5 % level of significance, the table value of t is 1.96.

The calculated value of t is 4.55 which is greater than the table value.

Hence the null hypothesis is rejected.

So there is significant difference between male students and female students in their social anxiety.

Illustration

Two samples of 100 electric bulbs each has a means 1000 and 1550, standard deviation 50 and 60. Can it be concluded that two brands differ significantly at 1% level of significance in equality.

Solution

Let us take the hypothesis that there is no significant difference in the mean life of the two makes of bulbs.

$$t = \frac{M_1 - M_2}{\sqrt{\frac{S_1^{\,2}}{N_1} + \frac{S_2^{\,2}}{N_2}}}$$

$$= \sqrt{\frac{(50)^2}{100} + \frac{(60)^2}{100}}$$

$$= \sqrt{25 + 36}$$

$$= 7.81$$

$$\frac{Difference}{S.E} = \frac{1550 - 1000}{7.81}$$

$$= 6.4$$

Since the difference is more than 2.58 S.E (1% level of significance) The hypothesis is rejected. Hence there is a significant difference in the mean life of the two brands of bulbs.

5.4.6. Standard Error of the difference between Two Standard Deviations

In the case of two large random samples, each drawn from a normally distributed population, the S.E of the difference between the standard deviations is given by;

$$S.E\,(\sigma_1 - \sigma_2) = \sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$$

Where populations standard deviations are not known.

$$S.E\,(S_1 - S_2) = \sqrt{\frac{S_1^2}{2n_1} + \frac{S_2^2}{2n_2}}$$

Illustration

In a sample of 1000 the mean is 17.5 and the s.d 2.5. In another sample of 800 the mean is 18 and standard deviation 2.7. Assuming that the samples are independent discuss whether the two samples can have come from a population which have the same standard deviation.

Solution: Let us take the hypothesis that there is no significant difference in the standard deviations of the two samples.

$$S.E\,(\sigma_1 - \sigma_2) = \sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$$

$\sigma^1 = 2.5,\ n_1 = 1000,\ \sigma^2 = 2.7,\ n_2 = 800$

$$S.E(\sigma_1 - \sigma_2) = \sqrt{\frac{(2.5)^2}{2000} + \frac{(2.7)^2}{1600}}$$

$$= \sqrt{\frac{6.25}{2000} + \frac{7.29}{1600}}$$

$$= \sqrt{0.00312 + 0.00456}$$

$$= 0.876$$

$$\frac{Difference}{S.E} = \frac{2.7 - 2.5}{0.0876}$$

$$= \frac{.2}{0.0876}$$

$$= 2.283$$

Since the difference is more than 1.96 S. E at 5% level of significance the hypothesis rejected. Hence the two samples have not come from a population which has the same standard deviation.

Hypothesis Testing of Correlation co-efficient

We may be interested in knowing whether the correlation co-efficient that we calculate on the basis of sample data is indicative of significant correlation. For this purpose we may use normally the t-test or the F-test depending upon the type of correlation co-efficient we use the following tests for the purpose.

a) In the case of simple correlation co-efficient we use t-test and calculate the test statistics as under.

$$t = r_{xy}\sqrt{\frac{n-2}{1-r_{xy}^2}}$$

with n-2 degrees of freedom r_{xy} being co-efficient of simple correlation between x and y.

This calculated value of 't' is then compared with its table value and if the calculated value is less than the table value, we accept the null hypothesis of the given level of significance. We may infer that there is no significant relationship of statistical significance between two variables.

b) In the case of partial correlation co-efficient. We use t-test and calculate the test statistics as under.

$$t = r_p \sqrt{\frac{n-k}{1-r_p^2}}$$

with n-k degrees of freedom, n being the number of paired observations and k being the number of variables involved, rp happens to be the co-efficient of partial correlation. If the value of 't' in the table is greater than the calculated value, we may accept the null hypothesis and infer that there is no correlation.

c) In the case of finding difference between the correlation of two sample; we may use z-test.

z-test was devised by prof. Fisher to test the significance of the correlation co-efficient in small samples. In this method, the co-efficient of correlation is transformed into z. This is used to test-

(i) Whether an observed value of r differs significantly from some hypothetical value,

(ii) Whether two sample values of r differ significantly in order to apply the test, values of z, η by appling fishers transformation and then calculate the value of the standard normal variance

$$\frac{Z-\eta}{1/\sqrt{n-3}}$$

$$Z = \frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right) \text{ or } 1.1513\log_{10}\left(\frac{1+r}{1-r}\right)$$

$$\eta = \frac{1}{2}\log_e\left(\frac{1+S}{1-S}\right) \text{ or } 1.1513\log_{10}\left(\frac{1+S}{1-S}\right)$$

Where S is population correlation co-efficient.

$$SE_z = \frac{1}{\sqrt{n-3}}$$

consider an example where r = 0.50 and N = 20. verify whether the correlation co-efficient significant at 5% level.

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

$$= 50\sqrt{\frac{20-2}{1-(50)^2}}$$

= 2.45

The df 20-2 = 18. At 5% level of significance for 18 df, the table value of 't' is 2.10. The calculated value of t is greater than the table value. Hence the null hypothesis is requested.

Test the significance of the correlation r = 0.5 from a sample of size 18 against hypothesis correlation S = 0.7.

Solution

To test the hypothesis that correlation in the population is 0.7.

Applying Z – transformation

$$Z = \frac{1}{2}\log_e \frac{1+r}{1-r}$$

$$= 1.1513 \log_{10} \frac{1+0.5}{1-0.5}$$

$$= 1.1513 \times 0.4771$$

=0.549

$$\eta = \frac{1}{2}\log_e \left(\frac{1+S}{1-S}\right)$$

$$= 1.1513 \log_{10} \frac{1+0.7}{1-0.7}$$

$$= 1.513 \log_{10} 5.67$$

$$= 1.513 \times 0.7536$$

$$\eta = 0.868$$

$$Z - \eta = 0.549 - 0.868$$

= 0.319

$$F_{az} = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{18-3}} = \frac{1}{\sqrt{15}} = 0.258$$

$$t = \frac{Difference}{SF_{az}} = \frac{0.319}{0.258} = 1.236$$

since the value is less than 1.96 at 5% level of significance, it could have arises due to fluctuations of sampling. Hence it can be said that $\delta = 0.7$

The correlation between intelligence scores and mathematics marks of two sections of IX standard students are given as follows.

	A	B
Sample size	5	12
Value of r	.870	0.560

Test the significance of the difference between two values using Fisher's Z transformation.

Solution

Let the hypothesis be that the sample are drawn from the some R population.

Applying z- test

$$Z = \frac{Z_1 - Z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

$$Z_1 = \frac{1}{2}\log_e \frac{1+r_1}{1-r_1}$$

$$= 1.513 \log_{10} \frac{1+0.83}{1-0.83}$$

= 1.3333

$$Z_2 = \frac{1}{2}\log_e \frac{1+r_2}{1-r_2}$$

$$= 1.513 \log_{10} \frac{1+0.56}{1-0.56}$$

= 0.633

Z_1-Z_2 = 1.3333 – 0.6333

= 0.7

$$SE_{21-22} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}$$

$$= \sqrt{\frac{1}{5-3} + \frac{1}{12-3}}$$

= 0.782

$$t = \frac{Z_1 - Z_2}{SE_Z}$$

$$= \frac{0.7}{0.782} = 0.895$$

Since the value is less than 1.96 at 5% level of significant, the hypothesis is accepted.

ANALYSIS OF VARIANCE

This method is devised by R.A. Fisher in 1923. It is also known as F – test, the F stands for fisher. This F – test is an improvement over 't' test. The 't' test is used for ascertaining the significance of difference of two means while F – test is used for testing the significance of difference more than two means simultaneously. The 't' test evaluates the significance of difference between variance only while F – test examines both between variance as well as within variance. The analysis of variance is associated with the design of experiments.

Meaning of the term 'Analysis of variance'

A composite procedure for testing simultaneously the difference between several sample means is known as the analysis of variance. The term 'analysis of variance' deals with the task of analyzing of breaking up the total variance of a large sample or a population consisting of a number of equal groups or sub-samples into two components (two kinds of variance), given as follows.

1. "Within-groups" variance. This is the average variance of the members of each group around their respective group means, i.e. the mean value of the scores in a sample (as members of each group may vary among themselves).
2. "Between-groups" Variance this represents the variance of group means around the total or grand mean of all groups. i.e the best estimate of the population mean (as the group means may vary considerably from each other).

ANOVA

When we want to compare more than two groups we can use analysis of variance.

Formula

$$F = \frac{MSVB}{MSVW}$$

where,

MSVB = Mean square variance between groups.

MSVW = Mean square variance within groups.

The significance of F ratio is determined from a Table. Table indicates the F critical values necessary to reject the null hypothesis at selected levels of significance.

5.5.1. Procedure for calculating the analysis of variance

Step 1: Arrangement of the given table and computation of some initial values. In this step, the following, values needed in computation are calculated from the experimental data arranged in proper tabular form:

i) Sum of squares, $\Sigma X_1, \Sigma X_2, \ldots\ldots$ and the grand sum, ΣX

ii) Group means, $\frac{\Sigma X_1}{N}, \frac{\Sigma X_1}{N}, \ldots\ldots\ldots$ and the grand mean $\frac{\Sigma X}{N}$

iii) Correlation term C computed by the formula

$$C = \frac{(\Sigma X)^2}{N} = \frac{Square\,of\ the\ grand\ sum}{Total\ No.of\ cases}$$

Step 2. Arrangement of the given table in to squared-form table and calculation of some other values. The given table is transformed into a squared-form table by squaring the values of each score given in the original table and then the following values are computed.

i) $\Sigma X^2_1, \Sigma X^2_2, \Sigma X^2_{3,\ldots\ldots}$

ii) ΣX^2

Step 3. Calculation of total sum of squares. The total sum of squares around the general mean is calculated with the help of the following formula:

$$S_1^2 = \Sigma X^2 - Correction\ value@$$

$$= \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$

Step 4 . Calculation of between-group sum of squares. The value of the between-groups sum of squares may be computed with the help of the following formula.

$$S_b^2 = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3} + ... \frac{(\Sigma X_n)^2}{n_n} - C$$

$$= \frac{(Sum\ of\ scores\ in\ group\ I)^2}{No.of\ scores\ in\ group\ I} + \frac{(Sum\ of\ scores\ in\ group\ II)^2}{No.of\ scores\ in\ group\ II} + ... - C$$

Step 5. Calculation of between-groups sum of squares. Between-groups and within-groups sum of squares constitute the total sum of squares. Therefore, after steps 3 and 4, the value of within groups sum of squares (sum of squares of deviation within each group about their respective group means) may be calculated by subtracting the value of between-groups sum of squares from the total sum of squares. Its formula, therefore, goes thus:

Within-groups sum of squares.

$$S_w^2 = S_t^2 - S_b^2$$

Step 6 : Calculation of the number of degrees of freedom. All these sums of squares calculated in steps 3-5 possess different degrees of freedom given by

Total sum of squares, (S_t^2) =N-1

Between-groups sum of squares, $(S_b^2) = K - 1$

Within-groups sum of squares, $(S_w^2) = (N-1) - (K-1) = N - K$

Where N represents the total number of observations, scores or frequencies and K, the number of groups in the research study.

Step 7. Calculation of F-ratio. The value of F-ratio furnishes a comprehensive or overall test of significance of the difference between means. For its computation, we have to arrange the data and computation work in the following manner:

Source of variance	Sum of squares	df	Mean square variance
Between-group	S_b^2 (*computed in step* 4)	K-1	$\frac{S_b^2}{K-1}$
Within-groups	S_w^2 (*computed in step* 5)	N-K	$\frac{S_w^2}{N-K}$

$$F = \frac{Mean\ square\ \text{var}\ iance\ between-groups}{Mean\ square\ \text{var}\ iance\ within-groups}$$

Formula

$$F = \frac{MS_b}{MS_w}$$

$$MS_b = \frac{SS_b}{df_b}; MS_w = \frac{SS_w}{df_w}$$

$$SS_b = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + ...\frac{(\Sigma X)^2}{N}$$

$$SS_t = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$

$$SS_w = SS_t - SS_b$$

Where

MS_b - Mean square between

MS_w - Mean square within

SS_b - Sum of square between

SS_w - Sum of squares within

df – Degrees of freedom

n – Number of subjects

$$df_w = N - K$$

$$df_b = K - 1$$

Step 8. Interpretation of F-ratio. F ratio are interpreted by the use of the critical value of F-ratios given in the ANOVA table. This table has the number of degrees of freedom for the greater mean square variance on the left-hand side. If our computed value of F is equal or to greater than the critical tabled value of F at a given level of significance 0.05 or 0.01, it is assumed to be significant and consequently, we reject the null hypothesis of no difference among these means at that level of significance. However, a significant F does not tell us which of the group means differ significantly; it merely tells us that at least one mean is relatively different from some other. Consequently, thee arises a need for further testing to determine which of the differences between mean are significant.

Incase our computed value of F is less than the critical tabled value of F at a given level of significance, it is taken as non-significant critical tabled value of F at a given level of significant, it is taken as non-significant and consequently, the null hypothesis cannot be rejected. Then there is no reason for further testing (as none of the differences between means will be significant). In a summarized form, the above analysis may be represented as follows:

F significant Null hypothesis rejected Need for further testing

F Non-significant Null hypothesis not rejected No need for further testing

Generally, as and when we get the value of F as less than I we straightaway interpret it as non-significant resulting in the non-rejection of the null hypothesis.

Step 9. Testing differences between means with the t test. When F is found significant, the need for further testing arises. We take pairs of the group means one by one for testing the significance of difference. The t test provides an adequate procedure for testing the significance when we have means of only two samples or groups at a time for consideration. Therefore, we make use of the t test the differences between pairs of means.

$$t = \frac{D}{\sigma_D} = \frac{Difference\ between\ two\ means}{S\tan dard\ error\ of\ the\ difference\ between\ the\ means}$$

and σ_D is computed by the formula

$$\sigma_D = \sigma\left[\frac{1}{n_1} + \frac{1}{n_2}\right]$$

Where

σ = Pooled SD of the samples drawn from the same population.

n_1, n_2 = Total No. of cases in samples I and II, respectively.

In the analysis of variance technique, within-groups means square variance provides us the value of σ^2, the square root of which can give us the required pooled SD of the samples of groups included in our study.

The degrees of freedom for within-group sum of squares are given by the formula N-K. With these degrees of freedom we can read the t values from Table C given in the Appendix, at the 0.05 and 0.01 levels of significance. If the computed value of t is found to be equal to or greater than the critical tabled value of t at 0.05 or 0.01 levels, we can reject the null hypothesis at the level of significance.

Proceeding similarly, we take other pairs for testing the difference between means and arrive at conclusions.

The following are the fundamental assumptions for the use of analysis of variance technique:

1. The dependent variable which is measured should be normally distributed in the population.
2. The individuals being observed should be distributed randomly in the groups.
3. Within-groups variances must be approximately equal.
4. The contributions to variance in the total sample must be additive.

Example

An experiment is conducted to study the effectiveness of 3 methods-lecture, question answer and library methods. In each group four students are assigned randomly. The obtained scores are given in the following table. Is there significant different among 3 methods of teaching?

Methods of Teaching

Lecture (X_1)	Question answer (X_2)	Library (X_3)
4	9	2
5	10	4
1	9	2
2	6	2

Solution

X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
4	16	9	81	2	4
5	25	10	100	4	16
1	1	9	81	2	4
2	4	6	36	2	4
$\sum X_1 = 12$	$\sum X_1^2 = 46$	$\sum X_2 = 34$	$\sum X_2^2 = 298$	$\sum X_3 = 10$	$\sum X_3^2 = 28$

$$\text{Grand Total } T = \sum X_1 + \sum X_2 + \sum X_3$$

$$= 12 + 34 + 10$$

$$\mathbf{T = 56}$$

Number of scores N = 12

Correction Factor (CF)

$$CF = \frac{T^2}{N}$$

$$= \frac{56^2}{12} = \frac{3136}{12}$$

CF = 261.33

Total Sum of square (SST)

$$SST = \left(\sum X_1^2 + \sum X_2^2 + \sum X_3^2\right) - CF$$

$$= (46 + 298 + 28) - 261.33$$

$$= 372 - 261.33$$

SST = 110.67

Sum of square between groups (SSB)

$$SSB = \left(\frac{\left(\sum X_1\right)^2}{n} + \frac{\left(\sum X_2\right)^2}{n} + \frac{\left(\sum X_3\right)^2}{n}\right) - CF$$

$$= \frac{(12)^2}{4} + \frac{(34)^2}{4} + \frac{(10)^2}{4} - 261.33$$

$$= \left(\frac{144}{4} + \frac{1156}{4} + \frac{100}{4}\right) - 261.33$$

$$= (36 + 289 + 25) - 261.33$$

$$SSB = 350 - 261.33$$

SSB = 88.67

Sum of square within groups (SSW)

$$SSW = SST - SSB$$

$$= 110.67 - 88.67$$

SSW = 22

Degrees of freedom (df)

Degree of freedom between group is **Number of groups – 1.**

df (Between) = (3 -1)

= 2

Degree of freedom within group is **Number of scores – number of groups.**

df (within) = (2 – 3)

= 9

ANALYSIS OF VARIANCE TABLE

Source of Variation	Sum of Square	df	Mean Square Variance	F
Between	88.67	(3-1) = 2	44.33	18.17
Within	22.00	(12-3) = 9	2.44	

$$MSVB = \frac{SSB}{df\ (Between)}$$

$$= \frac{88.67}{2}$$

MSVB = 44.34

$$MSVW = \frac{SSW}{df\ (within)}$$

$$= \frac{22}{9}$$

MSVW = 2.44

$$F = \frac{MSVB}{MSVW}$$

$$= \frac{44.33}{2.44}$$

F = 18.17

CHI-SQUARE TEST

The chi-square test is an important test among the several test of significance developed by statisticians. Chi-square, symbolically written as χ^2 (pronounced as ki-square) is a statistical measure used in the context of sampling analysis for comparing a variance to a theoretical variance. It is a non-parametric test. It can be used to make comparison between theoretical data and actual data of a population. This is useful for all

researcher to (i) test the goodness of fit; (ii) test the significance of association between two attributes, and (iii) test the homogeneity or the significance of population variance.

Chi-square can be defined as

$$\chi^2 = \Sigma\left(\frac{(O-E)}{E}\right)^2$$

When O is the observed frequency; E is the expected frequency.

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

Where RT is the row total for the row containing the cell

CT is the column total for the column containing cell

G_T is the total number of observations.

While interpreting the calculated value of χ^2 is compared with the taller value. For, this degrees of freedom are essentially calculated. Degrees of freedom means the number of classes to which the values can be assigned arbitrarily or it will without violating the restriction or limitation placed.

$$df = (r-1)(c-1)$$

Where r – No of rows and c – Number of columns

5.6.1. Conditions for the Application of χ^2 Test

The following conditions should be satisfied before χ^2 test can be applied:

(i) Observations recorded and used are collected on a random basis.

(ii) All the items in the sample must be independent.

(iii) 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.

(iv) 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.

(v) The constraints must be linear. Constraints which involve.

5.6.2. Steps involved in Applying Chi-square test

The various steps involved are as follow:

(i) 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 x 2 or any contingency table, the expected frequency for any given cell is worked out as under:

$$Expected\ \ frequency\ of\ any\ cell = \left\{ \frac{(Row\ total\ \ for\ the\ row\ of\ \ that\ cell) \times (Column\ total\ \ for\ the\ column\ of\ \ that\ cell)}{(Grand\ total)} \right\}$$

(ii) Obtain the difference between observed and expected frequencies and find out the squares of such differences i.e., calculate $(O_{ij} - E_{ij})^2$.

(iii) Divide the quantity $(O_{ij} - E_{ij})^2$ obtained as stated above by the corresponding expected frequency to get $(O_{ij} - E_{ij})^2$ / E_{ij} and this should be done for all the cell frequencies or the group frequencies.

(iv) Find the summation of $(O_{ij} - E_{ij})^2$ /E_{ij} values or what we call $\Sigma \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$. This is the required χ^2 value.

Problem

1. The opinion of boys & girls regarding homework are given.

	Negative	**Positive**	**Total**
Boys	20	80	100
Girls	40	60	100
Total	60	140	200

Is there any significant association between sex and attitude towards home work.

Null Hypothesis

There is no significant association between sex and attitude towards homework.

How to find the ‘E’ value?

$E = \frac{C_T \times R_T}{G_T}$ C_T - Column total

R_T -Row total

$E(20) = \frac{60 \times 100}{200} = 30$

$E(80) = \frac{140 \times 100}{200} = 70$

$E(40) = \frac{60 \times 100}{200} = 30$

$E(60) = \frac{140 \times 100}{200} = 70$

O	**E**	**O-E**	**(O-E)^2**	$\frac{(O-E)^2}{E}$
20	30	-10	100	3.33
80	70	10	100	1.43
40	30	10	100	3.33
60	70	-10	100	1.43
		Σ =0	400	9.52

$df = (r-1) \times (c-1)$ (without the total column & row)

$= (2-1) \times (2-1)$

= 1

At 5% level of significance for I df, the table value χ^2 is 3.841.

Conclusion

The calculated value of χ^2 is more than the table.

-we reject the null hypothesis. So, there is significant association between sex and attitude towards homework.

2. The following table gives the performance in skill of male and female.

Using chi-square, test whether there is any association between the sex and nature of work?

Sex	Nature of work		
	Skilled	Unskilled	Total
Male	40	20	60
Female	10	30	40
Total	50	50	100

Null Hypothesis

There is no association between the sex and nature of work.

$$E = \frac{R_T \times C_T}{G_T}$$

$$E(40) = \frac{60 \times 50}{100} = 30$$

$$E(20) = \frac{60 \times 50}{100} = 30$$

$$E(10) = \frac{50 \times 40}{100} = 20$$

$$E(30) = \frac{50 \times 40}{100} = 20$$

Calculate the value of chi-square value for the data

Agree	Undecided	Disagree	Total
15	24	6	45

Calculation of Chi-Square Value

	Agree	Undecided	Disagree	Total
O	15	24	6	45

E	15	15	15	
O-E	0	9	-9	
$(O-E)^2$	0	81	81	
$\frac{(O-E)^2}{E}$	0	5.4	5.4	$\frac{(O-E)^2}{E}$ **=10.8**

.

Calculate the value of chi-square value for the data

SA	A	UD	DA	SDA	Total
10	25	30	10	15	90

Calculation of Chi-Square Value

	SA	A	UD	DA	SDA	Total
O	10	25	30	10	15	90
E	18	18	18	18	18	
O-E	-8	7	12	-8	-3	
$(O-E)^2$	64	49	144	64	9	
$\frac{(O-E)^2}{E}$	3.00	2.72	8.00	3.60	0.5	

χ^2 = 18.30.

5.6.3. Advantages and limitations of chi-square test

Advantages

1. It is mainly a non-parametric test but it is used as parametric and non-parametric tests

2. It is applicable for the small sample and large sample
3. It is manly used for nominal scale, but it is used for interval and ordinal scales
4. It is used for finding the association of two variables in terms of expected and observed frequencies.
5. It is used to test the homogeneity of data.
6. It is a test of finding the goodness of fit of distribution.

Limitations

1. In this technique ,the main emphasis is given on the frequencies rather than scores, even when data on interval scale.
2. χ^2 test can be applicable to two variables but it is difficult to apply to more than two variables.
3. It is used for both inferential as well as descriptive purpose. Therfore, it is difficult to decide whether it is parametric or non-parametric.