Bachelor of Computer Application Part–III

Paper : BCA–306

Computer Oriented Numerical and Statistical Methods

Unit 1

Department of Distance Education
Punjabi University, Patiala

Lesson Nos. :

LESSON NO. 1.1 **AUTHOR : AMANPREET KAUR**

INTRODUCTION TO STATISTICS

1.1.1 Introduction

Growing complexities of human natural/social phenomenon has left no alternative to the world decision makers to depend upon the cordinal as well as ordinal methods of measurement and interpretation of the varied problems. Nearly, every activity-mental or physical or natural, is measured and interpreted quantitatively. Today whole world lives in the world of numbers/Statistic and reaches to conclusive decisions on the basis of statistical knowledge. Perhaps this may be the reason that Solomon Fabricant visualised the prospects of development of statistical use in these words. "Increasing public interest in and demand for social statistics rests on the basic premise that the problem of the society, as well as natural science and technology can be solved by the increase and diffusion of this especially matter-of-fact type of knowledge. The whole world now seems to hold that statistics can be useful in understanding, assessing and controlling the operations of Society." Also, the forecast H.G. tells that "statistical thinking will one day be as the ability, to read and write", seems to prove true. Today the authenticity of your statements is suspected if you have not expressed it numerically. "When you can measure what you are speaking about and express it in numbers, your knowledge is of a meagre and unsatisfactory kind". In short, statistical way of thought has entered the whole arena of human thinking in such a way that one becomes curious

to know about its origin and growth.

1.1.2 Meaning and Definition of Statistics

(a) *The Meaning :* Statistics, like many words have different meanings in different context. Some people regard statistics as data, facts or measurement while others believe it to be study of figures. There are another group of people who consider it as analysis of figures for forecasting or drawing inferences. Besides this, the representation of facts in the form of diagrams, a graph or maps is also supposed to be statistics. Processing, analysis and application of quantitative facts is regarded as statistics. In this way statistics includes three forms.

(i) *As Numerical Data :* In the product form, it represents the numerical data, such as statistics of national income, unemployment, imports and exports, etc. It is used here in plural form.

(ii) *As a Subject :* In the process form, it is used as subject and is in Singular form, like Economics, Physics etc. In this sense, the term Statistics' refers the whole field of study of which 'Statistics' in the plural sense are the subject-matter. In other words, it refers to the statistical principles and methods used in collection, analysis and interpretation of data. These methods finally help in taking decisions and testing the hypothesis.

(iii) *In its modern connotation :* It may also refer to the study and research into the theory and principle underlying statistical methods. It is the field of study that expands the frontiers of statistical methodology and uses.

(iv) "Statistics" is also used by the experts in the field, for the terms like mean, median, mode, standard deviations, etc. calculated from sample.

In brief statistics is used both in singular and plural form. It is also used to represent (i) Numerical data, (ii) Statistical science subject, and (iii) statistical measures, which is clear from Tate's this statement 'You compute statistics, by statistics from statistics."

1.1.3 Nature & Scope of Statistics

The Nature : As we have seen above that some statisticians have put statistics in the category of science, while others believe it to be arts. There are many members who believe it to the both Art and Science. To decide the nature of statistics it has to be examined in both the categories.

(a) *Statistics as a Science :* "Science is a body of systematized knowledge". In this way, any subject can be put in the category of science if it possess following characteristics :

(i) It is a systematized group of knowledge.

(ii) Its laws and methods must be universally acceptable.

(iii) It must analyse the cause-effect relationship.

(iv) It must possess the quality of estimation and forecasting.

1.1.4 The Function of Statistics

Statistics performs many functions useful to human beings. Robert W. Buges elaborates the functions of statistics in these words. "The fundamental aspect of statistics is to push back the domain of ignorance, prejudice, rule of thumb, arbitray and premature decisions, tradition and dogmatism and to increase the domain in which decisions are made and principles are made on the basis of analysed quantitative facts". The evergrowing popularity as a quantitative methods is because of the functions, which statistics performs :

(i) *Statistics provides definiteness to the facts :* Quantitative facts can easily be believed and trusted in comparison to abstract and qualitative facts. Statistics summarises the generalised facts and presents them in a definite form. Various characteristics pertaining to some phenomena, become easily understandable, if they are expressed in numbers. For instance, the index number technique expresses the complex variables into a form, which can easily be understood. It is easy to understand that prices index of consumer items has gone up by 10 per cent, instead of saying that prices are increasing leaps and bounds.

1.1.5 Simplifies the complexities

It is very difficult for an individual to understand and conclude from huge numerical data. Statistical methods try to present the great mass of complex data into simple and understandable form. For example, statistical technique like mean, median, variation, correlation, graphs and diagram etc. make the complex data intelligible and understandable in short period and in better way. W.L. King defining the function of statistical science rightly wrote, "It is for the purpose of simplyfying these unwidely masses of facts that statistical science is useful. It reduces them to numerical totals or average which maybe abstractly handled like any other more numbers. It draws pictures and diagram to illustrate general tandencies and thus in many ways adapts these group of ideas to the capacity of our intellects."

1.1.6 The Function of Statisticians

A statistician is a person who collects the data with the help of statistical techniques for some definite purpose of enquiry, analysis and interprets the facts as they are. He is in other words, practitioner of art and science of statistics. Rhode divides the functions of three parts. "In the first place he is concerned with the assembling of stastical data, in the second place with their analysis, and in the third place with the interpretation of the results of such an analysis."

For the sake of convenience, the functions of a statistician can be described in four categories :

(i) *The Observation :* This is the first and most important function of a statistician. In the beginning, he ponders over the objectives of research and plans the enquiry after deep thinking about the time-schedule, economic situation and available resources, the area and scope of research, time involved, level of precision and accuracy desire, and modalities of data collection. It is also his duty to specify and decide the manpower required, suited to his needs. All these functions, which seems to be primary and ordinary, requires a thorough skill and expertise based on wide observation experience. This should be planned with great care and confidence.

(ii) *The Collection :* The collection of data with pre-determined and preplanned method is second important function of a statistician. It is he, who decides the method of collection of data, whether it should be through enumeration or through estimation? He collects both primary and secondary data for the purpose. To test the authenticity and accuracy of collected data, he edits them also and presents them in tabular form for the purpose of analysis and interpretation.

(iii) *The Analysis :* The analysis work of statistician is very wide and cumbersome. He has to perform many works like classifying, serialising and bring them to a comparable form. After this, he has to compute different statistical parameters like average, denoation standard deviation, skewness, regression coefficient, correlation coefficient, coefficient of association, etc., for establishing relationship between them.

1.1.7 Limitations of Statistics

Even though statistics have served the mankinds in many ways and in many front from peace to war and is being utilised by almost every field of knowledge for its advancement and further researches, it is not free from shortcomings which restrict its scope and usefulness. It is always advisable to use it by keeping its limitations in mind. Newsholms cautions about its limitations in these words "it must be regarded as an instrument of research of great value, but having several limitations, which are not possible to overcome and as such they need our careful attention". Tippett has also suggested to be careful in the use of statistics in these words, "The application of statistical methods of investigations in the technological and indeed in any other field is based on assumptions, is subject to limitations and often leads to uncertain results." These limitations are :

(i) *Statistics fails to study qualitative phenomenon :* Science of statistics, as

discussed, deals with a set of numerical data, and can be applied to the study of only those phenomena, which can be expressed in numbers like qualitatively or in numbers. But besides this, those facts which are not measurable/expressed in numbers like beauty, honesty, appreciation, intelligence, health, eagerness, etc. can not be studied unless these qualities/virtues/attributes are reduced into precise quantitative terms. Prof. Horace Secrist wrote this in these words. "Some phenomena can not be quantitativley measured, honesty of resourcefulness, integrity, good will, all important in industry, as well as in life, generally are not susceptable of direct statistical measurement". However, some statistical techniques like analysis of attributes, scaling techniques, weightage technique etc. can be used to qualitative phenomena indirectly. If they are assigned certain numerical scores/weightage. For example efficiency and efficacy of Malaria Eradication Programme can be studied by the number of persons saved from death suffering from malaria. Intelligence of a person can be studied by the marks obtained in a particular examination. Liking and disliking can be studied with the help of Projective Technique. Impact of heredity can be studied by coefficient of association or coefficient of collignation. All these measurement techniques will be indirect aspect of the phenomena and they have to be expressed quantitatively. In this way, it limits the scope of applicapability of statistical techniques to the study of quatifiable variables only.

1.1.8 Distrust of Statistics

There can not be two opinions about the utility and applicability of statistics for human welfare. This would enhance in future provided it gives trustworthy, accurate and valid inferences. The credibility of statistics is being questioned day-by-day. Public distrust is increasing. In other words, public loses its belief, faith and confidence in the science of statistics and starts condemmining it. This distrust has cropped up owing to improper use of statistical tools by unscrupulous, irresponsible, inexperienced and dishonest persons having no expertise in the field. This has been expressed by many experts in the field in different ways. Some people call it, "as a tissue of falsehood", Mark Twain once remarked that "there are three degrees of comparison in lying, lies, clammed lies and statistics", and treats statistics as the superlative degree of lying. According to La Gordia, wrapped statistics is better than Hitler's Big Lie, it misleads, yet it can not by pinned on you." People say that an aunce of truth will

produce tons of statistics, or statistics are lies of the first order. It has been remarked that, there are black lies, white lies, multichromatic lies', statistics is a rainbow of lies.

However, the general conviction that "Statistics can prove any thing, or what statistics reveal in ordinary, but what they write is vital; etc. compels an ordinary person to regard a statistician as naive, incaustious and something of psendo-magician. These statements clearly indicate the extent to which the science of statistics had come to diserpute.

However, there are people who have contrary opinions. The power of science of statistics is great. It can prove anything. There are always tow aspects of an event. It depends upon you how you look. But one thing is clearly true, fault does not lie in data but in the technique how it is used.

1.1.9 Exercise Set

1. 'Statistical methods are dangerous tools in the hands of the inexperts'. Explain fully the significance of the above statement.
2. "Figures do not lie." "Statistics can prove anything'. Explain and Reconcile the two statements.
3. (a) What is statistics ? Discuss its scope and limitations. (b) Write an essay on : "Statistics in the service of Trade and Commerce."
4. What are the shortcomings of statistics ? Can these shortcomings be overcome ?
5. Describe with the help of suitable illustrations, the functions of statistics.
6. "The proper function of statistics, indeed, is to enlarge, individual experience". Comment on the above statement and also explain the functions of a statistician.
7. Explain how in modern age statistics can be treated as the science of human welfare.
8. Write an essay on "The role of the Statistician in contemporary society".
9. Write an essay on "Statistics in the service of State".
10. 'Statistics are the strawout of which like every other economist, have to make bricks'. (Marshall). Elucidate this statement and indicate the utility of Statistics in Economic Planning in India.
11. "Statistics arose from practical requirements of problems in various shapers and its importance is due to its uses in treating such problems". Discuss giving suitable example.
12. Planning on the basis of inadequate and inaccurate statistics is worse than no planning at all." (Third Five Year Plan). Explain this statement and discuss the importance of statistics in the planned economic

development of India.

13. "Planning without statistics is a ship without rudder and compass". In the light of this statement, explain the importance of statistics as an effective aid to national planning in India.
14. "Statistics plays an important part not only in the study of Economics and Commerce, but also in actual business". Explain fully.
15. What is statistics ? Explain the importance of statistical.
16. "A knowledge of statistics is like the knowledge of foreign language or of algebra. It may prove of use at any time, under any circumstances." Explain.

BIVARIATE DATA

1.2.0 Introduction (Need)

After a careful planning of a statistical investigation, the task of collection of data is initiated. Collection of data is the back-bone of statistical investigation. All the other statistical activities like editing, classification and tabulation, analysis and interpretation, etc., start after the data have been collected. It is the foundation stone of statistical investigation. The accuracy and authenticity of all other statistical investigation depend upon the precision and universality of method of collection of data and data itself. An accurate and ambiguous data may leads to inaccurate and unauthentic inference. Hence, due care and precaution should be taken in collection of data for reaching to a valid and statistically authentic result.

1.2.1 Primary and Secondary Data

The data can be categorised as "primary" and "secondary" as per its method of collection and sources.

1. Primary Data : It is collected by an investigator or agency for the first time for the purpose of statistical investigations. It is first hand information. According to Prof. H. Secrist, "...primary data meant those data which are original, that is, those in which little or no grouping has been made, the instance being recorded or intemized as encounted. They are essentially raw materials, "Primary data once collected and published becomes secondary data."
2. Secondary Data : The data (published or un-published), which have been collected and processed by some agency or person, and are used by other agency or person for their statistical purpose. For example, population data collected and published by Registrar General of India, in the form of Censuses, will be secondary data for any agency or person;

who utilizes it for their statistical purposes. According to M.M. Blair, "Secondary data are those already in existence and which have been collected for some other purpose than answering of the question at hand."

(a) Primary data are original and like raw materials for statistical inquiry, where as secondary data are like finished/constructed goods, because they have been processed and analysed earlier also.

(b) Primary data are collected by some agency or person by using the method of data collection, where as secondary data are already collected and processed by some person/agency and is ready for use.

(c) The collection of primary data requires a considerable amount of money, time and personals as whole plan of investigation is initiated, where as secondary data are less time consuming and cheaper as they are taken from published/unpublished material.

(d) Primary data fulfill the requirements of statistical investigations ojectives where as secondary data are not always according to the objectives of the users. That is the reason that they are used after a deep filtration and alterations. In other words, it has to be critically examined.

In fact, the difference between primary and secondary data is only quantity, i.e. a matter of degree or relativity, not that of nature. The same set of data may be secondary in the hands of one and primary in the hands of others. In general, the data are primary to the source who collects and processes them for the first time and are secondary for all other sources who later use such data. According to Prof H Secrist, "The distinction between primary to secondary data is largely one of degree data which are secondary in the hands of one party may be primary in the hands of another.
(iv) Choice between Primary & Secondary Data : After seeing the major differences between primary and secondary data, the option of choice of using them will be determined by the approriateness of the data to the objectives of the investigations. Which type of data will best suit to the statistical inquiry, will be determined by many factors like nature, objective and scope of the inquiry, time and money in the hands of the investigator or agency, the degree of precision required and the status of the agency (whether government, State or Central, or Corporation-or Private Institutions or an individual).

Remarks

1. It is advised that primary data should be used only where (i) Secondary data are authentically not available, (ii) where a virgin field of

investigation is to explored i.e. for micro studies, (iii) where it is neccessary to test the facts available in secondary sources of data.

2. On the contrary, while using secondary data, it is best to obtain the data from the primary sources as far as possible. It will be useful as it is possible to know/discuss about the terminology used, units employed, size of sample and sampling techniquies (if applied), method of data collection, and method of analysis, etc. applied in the secondary data. It is also help to know the limits of the data. Besides this, the errors of transcription (if any) can well be avoided.

1.2.2 Methods of Collection of Primary Data

Primary data can be collected by using following methods : (1) The Observation (2) The Experimentation (3) The Investigation/Enquiry.

(A) Direct Personal Investigation

(B) Indirect Oral Investigation.

(C) Information through correspondents.

(D) Schedules sent through investigators.

(E) Mailed questionnarie/schedules method.

1.2.2.1 The Observation : This method of data collection is a direct checking of records, where informations are automatically or manually inserted.

The method of observation involves curious attention of the sense organs of investigators/or observer as stated in these words, "They encompass the most casual, uncontrolled experiences as well as the most exact film records of laboratory experimentation. They depend upon the preferences, alterness, range and depth of knowledge of the observers and on his goal of study. They are important determining factor in the selection of the pattern of observation. There may be wide differentials in observation, depending upon the observer's attention and training, knowledge and awareness. All of us notice some things and fail to see others. Observation, therefore, is a purely subjective phenomenon. It is an important technique of collecting information and occupies a significant place in data collection techniques. It facilitates the collection of fairly reliable and valid data. This is most important technique, generally used in psychostatistical studies. But it can be used to other field of knowledge also. For example, in traffic censuses (in which automatic devices record the number of vehicles passing) or in surveys of unauthorised car driving (where car drivers are checked for their authorised car driving licenses) by authorities, or in checking the stocks of published books of an author on record and on stock. However, observation technique of data collection is relatively less expensive method yielding accurate results. The important component/determinant of observation technique are sensation, attention and perception.

(I) Superiority of Observation Technique over other Data Collection Techniques: Observation technique has an edge over the other techniques of data collection in many ways :

(a) Unlike other techniques, which depend entirely on the retrospective or anticipatory picture of one's own behaviours, the observational technique has the advantage of recording the possible behaviour of the people as and when it occurs. In this way, it provides the data and sketches the behaviour of the respondent fully free from after-thought influences and other stresses and strains. It yields data that pertain directly to a typical behavioural situation.

(b) Sometimes, it is possible to take for granted behaviour, for they are supposed to be a part of habits. In the observational technique, however, nothing is taken for granted unnoticed. There is nothing which cannot be translated into language.

(c) This technique is superior. It facilitates the collection of information about those subjects who are unable to translate their feelings in language; for example, infants, animals and birds.

(d) The participation of the respondent is not necessary.

(2) Demerits : Apart from its merits, the observational techniques suffer from some serious limitations.

(a) It is not always possible to be present at the time of occurrence of an event, the anticipation of the occurrence of an event may fail owing to unforeseen circumstances.

(b) It is practically difficult to apply the observational technique because of the duration of an event. For example, private and confidential behaviour cannot be observed.

(c) Observation cannot be qualified.

(d) It does not prove into the underlying reasons for the behaviour observed.

(3) Remarks : Observational techniques, besides their limitations, are very useful in the formulation of hypothesis, the exploration of new areas of research, the collection of supplementary, material useful in interpreting the findings of other techniques, the collection of data for descriptive as well as experimental studies, and the testing of hypothesis. It is very useful technique for behavioural studies.

1.2.2.2. The Experimentation : Like observation, experimentation is a technique of collection of data in some special field of knowledge. For example, the impact of agricultural inputs like seed, fertilizer, irrigation, technology, and land, etc., on the production of different crops, can well be studied through the experiments, specially,

designed for the purpose. One can easily conduct a market research survey for knowing the scope of its product by making certain experiments.

1.2.2.3. The Investigation/Enquiry : (A) Direct Personal Investigation :

(1) The Meaning : In this method, investigators come in direct contact with the respondents at their places and collects the informations by observations and conversations. Investigator (or agency) has to establish direct contact with the respondent (Informers) for making enquiry and soliciting informations. It is also called interview method. Thus, if one wants to study the wages of household women worker (maind servant), he/she has to contact maid servants working in some houses of a locality. This information will be first hand and original. The validity and authenticity of this data (information) will depend on the honesty and sincerity of the investigator as well as respondent.

However, the scope of application of this technique is limited to micro-level intensive studies rather than macro-level extensive studies. In other words, this method may yield desired level of accuracy if the field of enquiry is limited i.e. local, confined to a definite locality, region or area. It is suitable when information have to be kept confidential. The desired level of authentic information can be improved only by the investigators, who possess qualities like skilness, tactfulness, inability and accuracy.

(2) The Merits : The advantages of direct personal investigation (or interview) methods are :

(i) Accuracy and Reliability : As the data is collected by the investigator himself, it is liable to be accurate and reliable.

(ii) Study of Abstract and Intangible Phenomena : It is helpful in the study of such abstract and intangible phenomena as mental status and the attitude of the respondent.

(iii) Study of Past Events : The method is advantageous in presenting a real picture of past events because the person involved in those events can give authentic information to the interviewer.

(iv) Collection of Wider Information : The method is most appropriate and useful in the collection of data on complex, evolutionally laden subjects to a large extent. It ensures a greater member of usable returns than other methods do without discriminating caste, creed and colour.

(v) Study of Complex Phenomena : The method is very useful in transforming the nature of complex events into a simple form, which makes it easier for one to study the joint form and its constituents validity. Poverty, unemployment, bribery, corruption, etc. are caused by many factors. This can very easily be studied by giving a relative weitage to interview information.

(vi) Knowledge of Secret Experience and Events : A skilled and experience

interviewer can had most secret information about an event from a person of a group, which is not possible by any other method.

(vii) Multi-dimensional Study : It is a powerful and appropriate method to study the multi-dimensional aspects of a problem, the gestures of the respondents and the environment affecting their life. The interviews are often conducted at the resident of respondent. This also helps the interviewer not only to observe the events but to go deeper in to the pros and cons of the problem.

(viii) Feasibility of Objective Study : Objective informations can be easily collected by using the structured or interview schedules and rating scales in standardised circumstances.

(ix) Obtaining Information from Children & Illiterates : The method is very useful in collecting informations from children and illiterates. It offers an opportunity to the interviewer to observe the respondent's habits, facial expressions and gestures.

(x) Flexibility in Study : The method offers an opportunity to the interviewer to collect supplementary information about the informant's personal characteristics and environment which is often of great value in interpreting results. It is flexible approach, allowing an opportunity for fresh questions, or check questions, if need be. It is in this way superior to other techniques.

(xi) Reasibility of Replication in some Cases : This method, in some cases, can be repeated, which ensures the validity of information. It also offers scope for the evaluation of information by reading of the expression of respondent. Any type of ambiguity can be removed at once by the interviewer.

(xii) Study of Spontaneous Responses and Facility of Inter Stimulation : The interviewer may catch the informant off-guard and thus get the most spontaneous reactions, which lacks in other methods.

(xiii) Facility of Establishing Raport : As the interview is a mutually stimulating social interaction, there is greater scope for friendliness with the respondent. It minimises the possibility of the absence of a response to any question. Besides above, it can advantageously be used for collecting uniform and homogenous information.

(3) The Demerits : Though, the interview technique is the best technique for collecting information from the informants, it too is not free from criticism on the following grounds :

(i) Requires Heavy Expenditure and Time : In terms of cost, energy and time, this technique is fairly expensive, because of the vast coverage of samples scattered over a large area, the chance of the absence of contact of the samples with the interviewer, the expenditure incurred. Sometimes whole process is so boring that the interviewer loses patience.

(ii) Doubtful information : As the technique and instruments used in the investigation are neither rigid, nor objective and definite. In this way, authenticity of the informaion is dubious and doubtful.

(iii) Difference in Values : The quality of information is also affected by the social status determining factor of the attitudes and viewpoints of inter viewer and respondents. It also creates difficulty in establishing a rapport with respondent by interviewer and hence affecting the quality of information. For example, an investigator of poor social status hesitates to contact the rich man, or even may not be able to establish a proper rapport with him. On the contrary, he may not be given due weightage and no authentic information can be obtained.

(iv) Emotional, Unbalances and Bias : If the interview last longer, it is quite natural that investigators as well as informant both lose their patience, which may affect the authenticity of information, likewise if the nature of respondents is bias, no authentic data can be obtained.

(v) Difficulty in Precise Recording : It is unscientific to recall information in order to record it. Nor one should use chemical appliances for recording.

(vi) Dependency of the Interviewers : As the interviewer has to make suitable adjustments in regard to fixing time, place, and data by contacting the respondent, who actually dominates the entire show, the interview remains one sided business. Investigators have to depend upon the respondent. In absence of confidence in interviewer-being stranger to respondent, may give prejudices information.

(vii) Problem of Trained Personal : For obtaining authentic informations from respondent it is necessary that investigator should be well trained in the field, beside being tactful, skilled, diplomat and have proper knowledge of language of the respondent. Supervision, checking and re-checking of information should also be known by the investigator.

(viii) Erroneous System : If the informant is a proxy and is of opposite sex, the information will positively be erroneous because of psycho-social factors. A man/or woman cannot give correct information for a woman/or man.

(ix) Difficult to Eschew Unbiased Information : There are many things, which cannot be verbalized or told; but they can be acted upon or easily shown. Besides this, it is also applicable to limited area samples, the scope of study is very limited. In this way, it cannot throw light on the characteristics of universe.

Remarks :

(1) Though the interview technique is not free from fallacies, this method is nevertheless very effective when it is carefully used by a trained, educated and experienced interviewer. W.I. King writes about this method in these words; "This type of inquiry, while admirable because of additional accuracy due to

personal supervisiion, must not cover to narrow a field to be representative and is also liable to too large an injection of the personal element. The prejudices and the desires of the investigators become too often unconsciously woven in to the fabric of his conclusions."

(2) These limitations can be over come by following techniques :

(i) The questions to be asked must be relevant and consistent.

(ii) The way of asking the questions should be standardised.

(iii) Non-directive techniques should be employed.

(iv) Supplementing interviews with projective techniques, wherever necessary.

(3) The Pre-requisities of a successful interview are :

(i) a proper study design, (ii) a friendly atmosphere.

(iii) a positive attitude, and (iv) unbiased and objective mind.

(4) The success of the method greatly depends on the personal qualities of the interviewer.

(5) The success of an interview is determined by "the quantity and quality of the information obtained and this depends on whether the respondent is properly handled and persuaded to cooperate with the interviewer. This requires a technique of winning over the respondent so that he may be creating a willing and cooperative participant. Only then, the interviewer can get a true and valid response from the respondent. It has to be born in mind that interviewing is an art, which is to be learnt. No interview can be successfully conducted unless it is based on scientifically tested interview tools i.e., interview schedules or questionnaire."

(4) Typology of Interview : Interviews can be of two types;

(1) Structured and (2) Unstructured.

In the structured or standardised interview, the questions, their sequence, and their wordings are fixed. No answers to the questions are pre-conceived and classified. It is advantageous because the results can easily be transferred with quantitative form.

Unstructured or unstandardised interviews are more flexible and open. The interviewer is free to choose and change the wordings of the questions without changing no schedule is used. It is based on the theme that interviewer is well-versed with the subject, and have the capacity and capability to talk, discuss the topic with respondent and draw the relevant informations from him. It is advantageous to build up confidence in respondent and his involvement may yield desired informations. It can be of many types viz, focused, clinical, Non-directive, and Repetitive.

1.2.2.3. (B) Indirect Oral Investigations :

(1) The Meaning : In this method of data collection, interviewer interviews third person who is directly indirectly closer to the person about whom information is sought, but he is reluctant to give information or may hide the information and present a false picture about him. It is very difficult to get correct information from a gambler, or a drugadict or a smoker or a corrupt person as it will destroy his social images. For getting a correct picture about these persons, one can interview and record the replieas of (i) persons who possess full factual information about the problem/person, (ii) persons not biased, (iii) persons capable of expressing themselves, and (iv) persons not motivated to give colour to the facts. These persons who are interviewed are known as witnesses and their answers are recorded. This method is generally used by Enquiry Committees or Commission. This is an oral inquiry technique.

(2) Remarks : The success of this technique depends upon the personal qualities like face, courage, intelligence, the capacity and capability of understanding the psychological and instinctive reactions, etc. and technique of approach of the interviewer. Such informations requires great care and vigilance for its assessment and such data should not be taken at their face value. Due allowance must be given for the conscious and unconscious bias of the informant.

(3) The Merits :

1. Vast Coverage : It is possible to cover large number of respondent or vast area under investigation in this method.
2. Economy : It is possible to have better results within limited time, money and man power.
3. Unbiased : Generally, it is free from human bias, i.e. of interviewer as well as respondent.
4. Expert's View : If necessary, the expert's views and suggestions of the specialists on the given problem can be obtained in order to formulate and conduct the inquiry more effectively and efficiently.

(4) The Demerits :

(i) Indirect information : The result can be erroneous because information is obtained from other persons, not directly connected.

(ii) Biased : If withnesses are biased, they may give the information keeping their interests at top.

(iii) Trained Personal Required : It is not always possible that to have the information through a trained and skilled interviewer, having sufficient knowledge of the field and subject. In that case, information may be faulty. To obtain reliable and accurate result from this data, it is

necessary that (i) many persons, not only one persons, be contacted for information, (ii) the integrity and unboundedness of the withnesses should be prior confirmed, (iii) the witness must be a bold and well-informed person and have full detail of facts about the problem, (iv) a proper allowance about the behaviour of witness must be made.

1.2.2.3. (C) Information through Correspondents :

(1) The Meaning : Under this method, information is not collected formally by investigator or enumerators. It is regularly sent by the correspondents appointed by investigator in different parts of area of investigation. These correspondents or agents in different regions collect the information according to their experience, decisions, liking etc., and then sent it to the investigator. This technique is popularily adopted by newspapers, periodicals and government for its monthly, fortnightly reports. You are aware that certain regular feature and news columns like sports, business trends, economic scenario etc. are regularly required by the papers and magazines. This is the method for their information. A more refined and sophisticated way of the use of this technique is the registration method in which any event like birth, death, is to be reported to the proper authorities, as and when or immediately after if occurs.

(2) The Merits :

(i) Economic Way : This method is very economic method so far as money and time is concerned. It covers a vast area through part time paid correspondence, who are paid according to the news items etc.

(ii) Vast Coverage : Information can be collected for vast region by this method.

(iii) Expediency : The required information can be obtained expeditiously, as rough estimates are required.

(3) The Demerits :

(i) Lack of Precision and Originality : The informations collected under this method lack orginality and precision, as they are estimates only.

(ii) Heterogeneous : The information is not homogeneous as it is collected by several correspondents using their own techniques, languages and estimates. This may cause the problem of informity and thus chance of less reliability.

(iii) Biasdness : If majority of correspondents are of one/some opinion, this way cause biasdness.

(iv) Test of Reliability : of the information is difficult in this method.

(v) Duration of Time : Sometimes, it is very difficult to have the news/ information within a desired time, owing to which the whole information becomes meaningless.

1.2.2.3. (D) Schedules Sent Through Invigilators :

(1) The Meaning : Under this method of data collection, investigators themselves are required to fill the informations in the schedules obtained from the respondents. The enumerators visit the places of house of the respondents and record the informations after enquiring about the facts. In this way the problems of reliability, incompleteness, inadequacy and non-responding, etc. are automatically solved in this method. These enumerators are well-trained in schedule filling and well-versed with the subject matter and notations and symbols used in the schedules. Enumerators are required to establish rapport with the respondents for which he must not act/do or show any gesture, which can create apathy or indifference or fear of leakage of secrecy or action in futures etc., and informations are hidden. As far as possible (always), no signature of the respondents should be taken anywhere. A cordial relationship is most suitable to get the correct informations. He should be well-aware about the customs and way of life of the people of the region. It is also necessary to have supervision over the investigators. This method is generally used by research institutions like N.C.A.E.R., I.I.P.S., E.I.C.C.I., etc., big business houses and large public enterprises, government agencies like Registrar General of India for series, C.S.O. for N.S.S. etc. where high degree of response and reliability is desired.

(2) The Merits :

(i) Wide Range : This technique of data collection can be used for collecting informations from large number of people coming from vast areas. For example, census operations, world fertility surveys by U.N.O. It can be used for extensive area.

(ii) High Level of Reliability and Precision : As the enumerators, who are trained and well-versed with the concepts and definitions of the symbols and terminologies, themselves record the informations. The reliability and precision of the data and result will be of higher order.

(iii) Personal or Face-to-Face Contact : As the ensumerators are face-to-face with the respondent, it is possible to explain the objective and help the respondents to understand the complex words and definitions. It will certainly lead to create a harmonious relationship, which will be helpful in getting reliable data from respondents.

(iv) Unbiasedness : There are little chances of biasedness as the enumerators are from both sides. Some of them like the subject and others dislike. Hence, a balance results.

(v) Verification of Facts : The results and the collected informations can be

checked and rechecked for its authenticity and validity. For example, in the recent census of 1991, Second round enumeration was conducted for checking and rechecking of the facts already collected.

(3) The Demerits :

(i) Expensive : It is fairly expensive method since it involves huge army of enumerators who are paid. The processing and handing of data itself is a costly affair.

(ii) Time Consuming : Such type of data collection techniques are more time consuming in comparison to other methods.

(iii) Supervision and Checking Difficulty : A large number of enumerators are engaged in the work of enumeration. There can be vast variation in the informations owing to personality differentials and quite a large number of supervisory staff is required to check the irresponsibility and carelessness in handling the data.

(iv) Greater Skillness : The success of the method and reliability of data requires greater skill among enumerators, which sometimes lacks.

(v) Well-defined Schedule : The success of getting correct informations from respondents very much depends upon the quality of the schedule, which requires higher skillness for its preparation.

1.2.2.3. (E) Mailed Questionnaire Method :

(1) The Meaning : This method is also known as 'Questionnaires to be filled by the informants method. Under this method, investigator prepares a questionnaire/schedule relating the object of inquiry, and sends them to individual informants with a request letter to return the questionnaire in the self-addressed (possibly stamped) envelope. The investigator also assures about maintaining secrecy about the information given by him. With questionnaire, he also educates the informants about the objective of the information/study and requests the informants to send them back within a prescribed period of time (a fixed date). This method is generally used by research scholars, private personnel, non-official agencies, and sometimes even the governing.

(2) Remarks :

(i) The success of this technique depends upon how the questions are replied and then returned ? It requires a special skill of framing the questionnaire and letter of request.

(ii) However, a supplementary letter explainting the terms in questionnaire should also be sent alongwith questionnaire.

(3) The Merits :

(i) Economical : Compare to other methods of data collection, it is the

cheapest one as one can collect informations covering a vast area within limited period and with only postal costs (which 10 always small). It also saves huge man power as required in other methods.

(ii) Originality : The information is given by the informants according to his choice, as there is no pressure of the investigators but moral.

(iii) Unbiased : Information is unbiased as he is not in face-to-face to investigators.

(iv) Reliable : This method is free from investigators interference/bias, hence comparatively reliable.

The chances of playing and trying with the information by investigator is meagre and hence less erroneous.

(4) The Demerits :

(i) Inadequate and Incomplete Informations : Most informants do not return questionnaire. Generally the non-respondents percentage range between 60 to 80. Not only this, these 20% who responds the questionnaire, gives the reply carelessly with an indifferent attitude and vague, informations incomplete and haphazard, which do not serve the purpose. It is rather inadequate and incomplete. This is a serious drawback of this method so far as reliability is concerned.

(ii) Less Accuracy and Precision : Quite often respondents suppress correct information and furnish wrong replies. It is not possible to verify the accuracy and reliability of the information received. In general, method suffers from low degree of reliability and precision.

(iii) Lack of Flexibility : In case of inadequate and incomplete answers, it is difficult to ask supplementary or complementary questions, hence, method suffers from inflexibility.

(iv) Limited Scope : This method is applicable only where respondents are educated.

(v) Possibility of getting Wrong Answers : There is a possibility of getting wrong results from the information received. In the words of Parton, "In a Public opinion poll, utilizing mailed questionnaire, a disproportionate high percentage of returns tend to come from people with extreme and/ or strongly felt opinion." Similarly, D. Gregogy and H. Ward expresses this in these words, "This lack of responser may bad to bias, called non-response bias", i.e. our result may not include a certain type of person with whom we wish to make contact and from whom we want information. Likewise, A.B. Blankenship wrote that, "the difficulty with partial response is that these who do answer the questionnaire are usually not representative of the entire group of the which the forms

have been sent."

1.2.2.4. Criterion of Selection of A Suitable Method :

It is not possible to say/point out a method from the primary data collection techniques, which will be suiting best in all the situations. None of them is faultless and possess universal applicability. In this light, one has to decide a method best suited to the objective and scope of an inquiry. There cannot be any definite rule for deciding this diploma. However, in selecting a method, following criterion factors can be taken into considerations :

(a) **The Nature of Inquiry :** The nature of the investigation plays important role is deciding the method of data collection best suited. If the nature of investigation is such that it is essential to establish a personal contact with the respondents, "The Direct Personal Investigation Technique" will be most appropriate, for example, the illiterate household woman labourers. For getting informations from educated people or institutions, "Mailing the questionnaire to Respondents Methods", will be appropriate. On the contrary, if the area of investigation is large and the objective/scope of coverage of subject is wide, it is better to collect the data through engaging large number of enumerators and 'Scheduled through Enumerator Method' will be best suitable. In conducting census of people, products, cattles, etc, this method is appropriate.

(b) **Object and the Scope of the Investigations :** The object and the scope of the investigation is another criterion, which determines the suitability of a method. For a limited area and wide ranging topics of confidential nature, direct contact method will be most appropriate. For sending regular and general nature of information for papers, magazines, etc., information through correspondents/ agents method will be most appropriate. Indirect oral investigation technique may yield good result in such cases.

(c) **Financial Constraints :** The financial resources available to the enumerators or the project is also a criterion on which method of data collection depends. The direct personal contact method requires huge financial resources compare to other methods.

(d) **Time Constraints :** The method is also determined by the time in hand for getting the information. If time is sufficiently large, then all other methods excepting through correspondents will be suitable.

(e) **Degree of Accuracy :** The another critrion which plays vital role in determining the method is the desired degree of accuracy. Where higher degree of accuracy of result is required in limited field of enquiry, direct personal investigation method will be most suitable. The informations obtained by indirect method possess less degree of accuracy. On the other

hand, informations through correspondents/agents are always estimates. The informations gather by postal method is neither complete, nor adequate. It is less accurate also.

All these criterion if taken into considerations will help in selecting a suitable and appropriate method of data collection. Besides this, the successes of getting a suitable degree of accuracy and reliability from the data collected by a particular method also depends upon the quality and approachability of investigators, as Prof. A.L. Bowley has remarked.In collection, "Common sense is the chief requisite, and experience the chief teacher."

LESSON NO. 1.3 AUTHOR : AMANPREET KAUR

CORRELATION ANALYSIS

1.3.0 Introduction

The measures of the central tendency and skewness throw light on the construction of a series. These measures are also used for comparison between two series of the same variable, but this is not enough sometimes. The term correlation indicates the relationship between two such variables in which changes in the values of one variable, the values of the other variable also change. Thus, correlation is a statistical technique which shows the relationship between two or more variables.

According to L.R. Conner, "If two or more quantities in sympathy so that movement in the one tend to be accompanied by corresponding movements in the other, they are said to be correlated."

According to Ya Lux Chou, "Correlation analysis attempts to determine the degree of relationship between variables."

We can see the relationship between various pairs of variables like, age and blindness, income and consumption, height and weight etc. In correlation, thus we deal with bivariate distributions.

1.3.1 Types of Correlation :

Correlation is described or classified in several different ways. It can be classified into :

(i) Positive or Negative Correlation :

On the basis of the direction of the change in the two variables, correlation can be-ve or +ve. If the change in both the variables is in the same direction i.e., if both increase simultaneously or decrease simultaneously, the correlation is said to be

positive.

If the change in both the variables is in the opposite direction i.e. if one increases, other decreases, then correlation is said to negative. The following examples would illustrate the difference between positive and negative correlation.

1. Positive Correlation :

Price (Rs.)	:	10	11	12	13
Supply	:	100	130	140	160

2. Negative Correlation :

Price (Rs.)	:	10	11	12	13
Demand	:	100	90	85	82

(ii) Simple and Multiple Correlation :

If we study, correlation between two variables, it is called simple correlation. In case, we study relation in more than two variables, it is called multiple correlation.

(iii) Linear and Non-Linear Correlation :

On the basis of the ratio of change in the related variables, the correlation can be linear or non-linear.

If the amount of change in a variable is at a constant ratio to the change in the other variable, the correlation is said to be linear. This relationship is represented by the equation $y = a + bx$ and when plotted on a graph paper, straight line is formed. This type of correlation is found only in physical sciences.

Illustration :

Price (Rs.)	:	10	11	12	13
Quantity Supplied	:	100	150	200	250

If the amount of change in one variable is not at constant ratio to the change in the other variable, the correlation is said to be non-linear. This type of correlation is generally found in social sciences.

Illustration :

Income (Rs.)	:	100	150	200	250
Consumption (Rs.)	:	70	100	120	130

Correlation is non-linear because for every increase in income by Rs. 50, the consumption firstly increases by Rs. 30, then by Rs. 20 and then by Rs. 10.

Properties of Correlation :

(i) Degree of correlation is indicated by coefficient of correlation. According to Prof. Karl Pearson, the coefficient of correlation 'r' varies between two limits i.e. ±1 i.e. the maximum value of coefficient of correlation can be + 1 and the minimum value of coefficient of correlation can be –1. It means correlation lies between –1 and +1 i.e. $-1 \leq r \leq +1$.

(ii) Correlation is not dependent on origin and change of scale. That means that there is no difference between step-deviation and simple deviation

methods.

1.3.2 Methods of Measuring Correlation :

Following methods can be used to measure the correlation between two variables :

1. Scatter diagram method.
2. Graphic method.
3. Karl Pearson's method.
4. Concurrent deviation method.

Here, the scattar diagram and graphic methods are not free from drawbacks. These methods provide us only the direction and amount of correlation between two variables. We describe the Karl Pearson's method for finding the Pearson's correlation coefficient 'r'.

1.3.3 Karl Pearson's Method

(a) Direct Method: Given by

$$r = \frac{\sum xy}{N\sigma_x\sigma_y}$$

The formula for Pearson's Correlation coefficient is.

$x = X - \bar{X}$, $y = Y - \bar{Y}$

σ_x = S.D. of x-series $= \sqrt{\frac{\Sigma x^2}{N}}$

σ_y = S.D. of y-series $= \sqrt{\frac{\Sigma y^2}{N}}$

N = Number of pairs of observations.

Covariance of x and $y = \frac{\Sigma xy}{N}$

So, we can write, $r = \frac{\Sigma xy}{N\sqrt{\frac{\Sigma x^2}{N} \cdot \frac{\Sigma y^2}{N}}}$

Thus $'r' = \frac{\Sigma xy}{N.\sigma_x\sigma_y}$

or $r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$

Example 8.1 : Calculate Karl Person's Coefficient of correlation from the following data

Income (Rs.)	:	10	12	18	24	23	27
Consumption (Rs.)	:	13	18	12	25	30	10

Solution :

Calculation of coefficient 'r'

Income			**Consumption**			
(X)	$x = X - \bar{X}$	x^2	(Y)	$y = Y - \bar{Y}$	y^2	xy
	X - 19					
10	-9	81	13	-5	25	45
12	-7	49	18	0	0	0
18	-1	1	12	-6	36	+6
24	+5	25	25	+7	49	+35
23	+4	16	30	+12	144	+48
27	+8	64	10	-8	64	-64
$\Sigma X = 114$	$\Sigma x = 0$	$\Sigma x^2 = 236$	$\Sigma Y = 108$	$\Sigma y = 0$	$\Sigma y^2 = 318$	$\Sigma xy = +70$
n = 6						

Mean of variable $\bar{X} = \frac{\Sigma X}{N} = \frac{114}{6} = 19$

Mean of variable $\bar{Y} = \frac{\Sigma Y}{N} = \frac{108}{6} = 18$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

$$\frac{70}{\sqrt{236 \times 318}} = \frac{70}{\sqrt{75046}} = \frac{70}{\sqrt{273.95}} = 0.255$$

(b) Short Cut Method :

In direct method, the mean is a whole number. Therefore, method is simple. But where mean is in fraction, it will involve difficult calculations. To solve is such a situation, short cut method is used. In this method, the deviations are calculated from assumed mean and following formula is used to calculate the coefficient of correlation (r) :

$$r = \frac{\Sigma d_x d_y - \frac{(\Sigma d_x \Sigma d_y)}{N}}{\sqrt{\Sigma d_x^2 - \frac{(\Sigma d_x)^2}{N}} \sqrt{\Sigma d_y^2 - \frac{(\Sigma d_y)^2}{N}}}$$

Where,

Σd_x = Sum of deviation of X series from its assumed mean i.e. $\Sigma (X - A)$

Σd_y = Sum of deviation of Y series from its assumed mean i.e. $\Sigma (Y - A)$

Σd_x^2 = Sum of square of deviation of X from assumed mean i.e. $\Sigma (X - A)^2$

Σd_y^2 = Sum of square of deviation of Y from assumed mean i.e. $\Sigma (Y - A)^2$

$\Sigma d_x d_y$ = Sum of Products of deviation of X and Y series from their respective assumed means $d_x.d_y = \Sigma (X - A) (Y - A)$

Example 8.2 :

Calculate coefficient of correlation from the following data :

Experience : (X)	16	12	18	4	3	10	5	12
Performance : (Y)	23	22	24	17	19	20	18	21

Solutions :

	A = 10			**A = 20**		
X	**$d_x = X - 10$**	**d_x^2**	**Y**	**$d_y = y - 20$**	**d_y^2**	**$d_x d_y$**
16	+6	36	23	+3	9	18
12	+2	4	22	+2	4	+4
18	+8	64	24	+4	16	+32
4	-6	36	17	-3	9	+18
3	-7	49	19	-1	1	+7
10	0	0	20	0	0	0
5	-5	25	18	-2	4	+10
12	+2	4	21	+1	1	+2
n = 8	$\Sigma d_x = 0$	$\Sigma d_x^2 = 218$		$\Sigma d_y = +4$	$\Sigma d_y^2 = 44$	$\Sigma d_x d_y = +91$

$$r = \frac{\Sigma d_x . d_y - \frac{\Sigma d_x . \Sigma d_y}{N}}{\sqrt{\Sigma d_x^2 - \frac{(\Sigma dx)^2}{N}} \sqrt{\Sigma d_y^2 - \frac{(\Sigma d_y)^2}{N}}}$$

$$r = \frac{91 - \frac{0 \times 4}{8}}{\sqrt{218 - \frac{(0)^2}{8}} \sqrt{44 - \frac{(4)^2}{8}}}$$

$$r = \frac{91}{\sqrt{9156}} \frac{91}{95.687} = 0.951$$

1.3.4 Rank Correlation :

Prof. C.E. Spearman has given a method of judging correlation between two attributes which cannot be measured in quantitative terms such as beauty wisdom, honesty, intelligence, etc. In other words, it is used when data are on qualitative nature. In such cases coefficient of correlation is calculated by using the following formula :

$$r_k = 1 - \frac{6\Sigma D^2}{n(n^2 - 1)}$$

Where r_k = Coefficient of rank correlation.

n = number of pairs of items.

ΣD^2 = sum of squares of differences in Ranks.

Core.I. When Ranks are given :

We have the following three cases :

(i) Take the differences of the two ranks i.e. $(R_1 - R_2)$ and denote these differences by D.

(ii) Square these differences and obtain ΣD^2

(iii) Apply the formula

$$r_k = 1 - \frac{6\,\Sigma D^2}{n\left(n^2 - 1\right)}$$

Example 8.3 : Calculate rank correlation coefficient between the ranks given for X and Y variables :

X	:	2	1	4	3	5	7	6
Y	:	1	3	2	4	5	6	7

X (R_1)	Y (R_2)	$(R_1 - R_2)$ D	D^2
2	1	1	1
1	3	-2	4
4	2	+2	4
3	4	-1	1
5	5	0	0
7	6	+1	1
6	7	-1	1
			$\Sigma D^2 = 12$

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

Here, $\Sigma D^2 = 12$ $\quad$ $N = 7$

$$r_k = 1 - \frac{6 \times 12}{7^3 - 7} = 1 - \frac{72}{343 - 7}$$

$$= 1 - \frac{72}{336}$$

$$= \frac{336 - 72}{336} = \frac{264}{336}$$

$r_k = 0.785$

Core. II. When Ranks are not given :

When we are given the actual data and not the ranks, it will be necessary to assign the ranks. Ranks can be assigned by taking either highest value as 1 on the lowest value as 1. The same method is used in case of both the variables.

Example 8.4 : Calculate the coefficient of correlation from the following data by Spearman's Rank difference method :

Price of Sugar (Rs.)	**Price of tea (Rs.)**
75	120
60	150
80	115
81	110
50	140

Solutions :

Price of Sugar (Rs.)	R_1	**Price of Tea (Rs.)**	R_2	$(R_1-R_2)^2 = D^2$
75	3	120	3	$0^2=0$
60	2	150	5	$(-3)^2=9$
80	4	115	2	$2^2 = 4$
81	5	110	1	$4^2 = 16$
50	1	140	4	$(-3)^2 =9$
				$\Sigma D^2 = 38$

$$r_k = 1 - \frac{6\,\Sigma D^2}{N\left(N^2-1\right)}$$

$$r_k = 1 - \frac{6\times 38}{120} = 1 = \frac{228}{120} = \frac{120-228}{120}$$

$$r_k = \frac{-108}{120} = -0.9$$

Core.III. When Ranks are Equal :

In such a case, it is necessary to give each individual an average rank. Thus, if two individuals are ranked equal at fifth and sixth place, they are each given the rank $\frac{5+6}{2}$ i.e. 5.5 while if these are ranked equal at fifth, sixth and seventh places, they are given the rank.

$$\frac{5+6+7}{3} = 6$$

Thus, when equal ranks are assigned to some entires, an adjustment in the above formula for calculating the rank coefficient is made.

$$r_k = 1 - \frac{6\left\{\Sigma D^2 + \frac{1}{12}\left(m^3 - m\right) + \frac{1}{12}\left(m^3 - m\right) + ...\right\}}{N^3 - N}$$

Where m = no. of items whose ranks are equal.

Example 8.5 :

Compute Speaman's Rank Correlation from the following data.

Marks in Economics :	50	60	65	70	75	40	70	80
Marks in Maths :	80	71	60	75	91	82	70	50

Solution :

Let the marks in Eco. be denoted as X and marks in Maths be denoted as Y

X	R_1	Y	R_2	R_1-R_2=D	$(R_1-R_2)^2 = D^2$
50	2	80	6	-4	16
60	3	71	4	-1	1
65	4	60	2	+2	4
70	5.5	7.5	5	+0.5	0.25
75	7	91	8	-1	1
40	1	82	7	-6	36
70	5.5	50	3	+2.5	6.25
80	8	50	1	+7	49
					$\Sigma D^2 = 113.5$

$$r_k = 1 - \frac{6\left\{\Sigma D^2 + \frac{1}{12}\left(m^3 - m\right) + \frac{1}{12}\left(m^3 - m\right) + ...\right\}}{N^3 - N}$$

$$= 1 - \frac{6\left[113.5 + \frac{1}{12}\left(2^3 - 2\right)\right]}{8^3 - 8}$$

$$= 1 - \frac{6\left(113.5 + 0.5\right)}{504}$$

$$= 1 - \frac{6 \times 114}{504}$$

$$= 1 - \frac{114}{84} = \frac{84 - 114}{84} \qquad r = -\frac{30}{84}$$

LESSON NO. 1.4 **AUTHOR : AMANPREET KAUR**

REGRESSION ANALYSIS

1.4.0 Introduction

Regression is a statistical device for measuring or estimating relationship between variables. Regression means to revert or to return back. The term was first introduced by Sir Francis Galton in 1877. He found, in his study of the relationship between the heights of fathers and sons, that tall fathers were likely to have tall sons and short fathers were likely to have short sons. However, the mean height of the sons of tall fathers was lower than the mean height of their fathers, and the mean height of the sons of short fathers was higher than the mean height of their short fathers. He referred to this tendency to return to the mean height of all men as regression, in his research paper.

In regression analysis, we have to assume one variable as the independent variable and the other as dependent variable. After estimating the relationship between variables, one can predict the most likely values of dependent variables, on the basis of given values of independent variables. Thus, it indicates the average relationship between two or more variables.

According to Ya-hum-Chau, "Regression analysis attempts to establish the nature of relationship between variables that is to study the functional relationship between the variables and thereby provide a mechanism for prediction or forecasting."

1.4.1 Difference between Correlation and Regression :

1. The correlation coefficient is a measure of degree of covariablity between two variables while the regression establishes a functional relationship between dependent and independent variables so that the former can

be predicted for a given value of the later.

2. Correlation merely ascertains the direction and degree of relationship between two variables, but it does not clearly specify as which variable is the cause and which is the effect. But this cause and effect relationship is clearly indicated by regression analysis.

1.4.2 Regression Lines :

The device used for estimating the value of one variable from the value of the other consists of a line through the points drawn in such a manner so as to represent the average relationship between the two variables. Such a line is called line of regression. There are two regression lines. One line is the regression of X on Y and the other is the regression of Y on X, described as :

I. Regression line of X on Y : The regression line of X on Y is formed by taking the most probable value of X for the given value of Y.

II. Regression line of Y on X : The regression line of Y on X is formed by taking the most probable value of Y for the given value of X.

Remarks : (i) These two regression lines show the average relationship between two variables. If there is perfect correlation (i.e. r = ±1), both the lines will coincide i.e. there will be only one line (see Fig. 1, Fig. 2.).

(ii) In case r = 0, both the lines will cut each other at right angle i.e. parallel to X-axis and Y-axis (Fig. 3).

(iii) These lines cut each other at the point of means of X and Y (see Fig 1).

(iv) Nearer these lines are, greater will be extent of correlation between X and Y, (Fig. 5, Fig. 6).

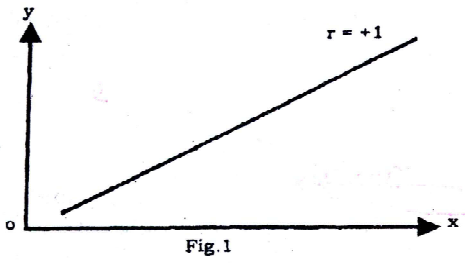


Fig.1

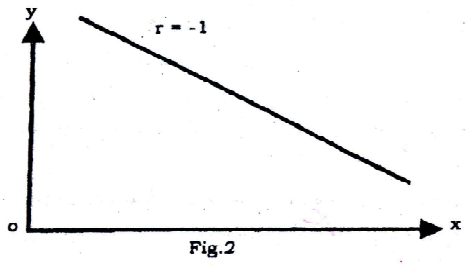


Fig.2

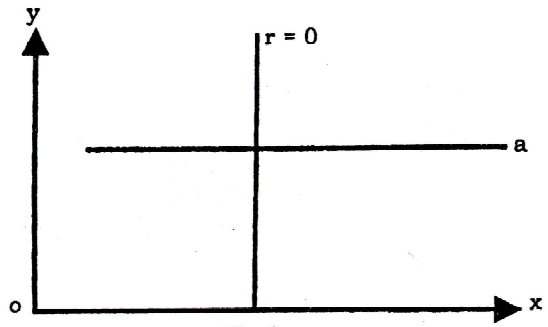


Fig.3

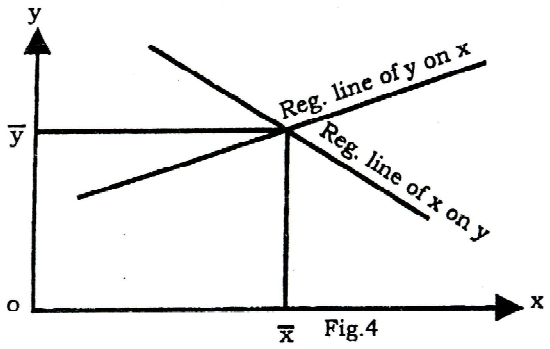


Fig.4

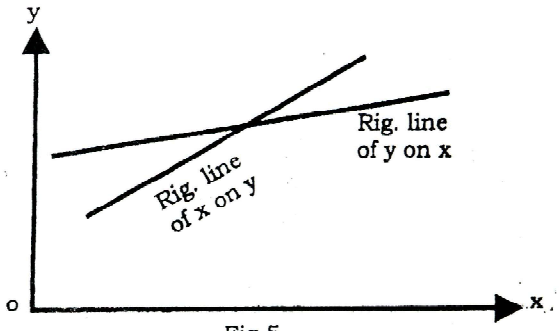


Fig.5

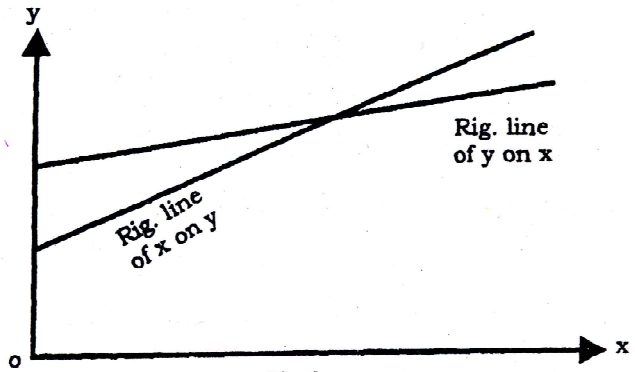
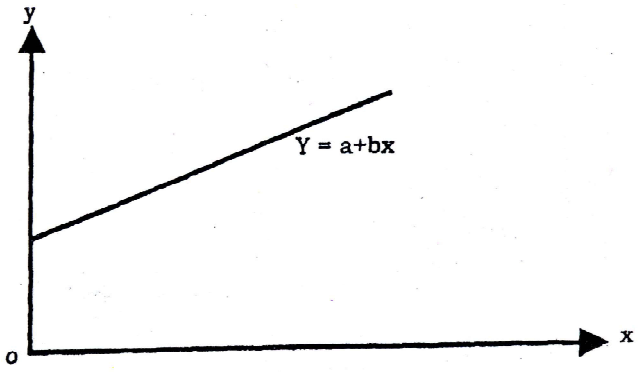


Fig.6

Assumptions : The regression lines are drawn on least square assumption which states that the sum of squares of the deviations of the observed 'Y' values from the fitted lines shall be minimum. The devitations from the points to the line of best fit can be measured in two ways - vertical, i.e. parallel to Y-axis and horizontal, i.e. parallel to X-axis. For minimizing the total of the squares separately, it is essential to have two regression lines.

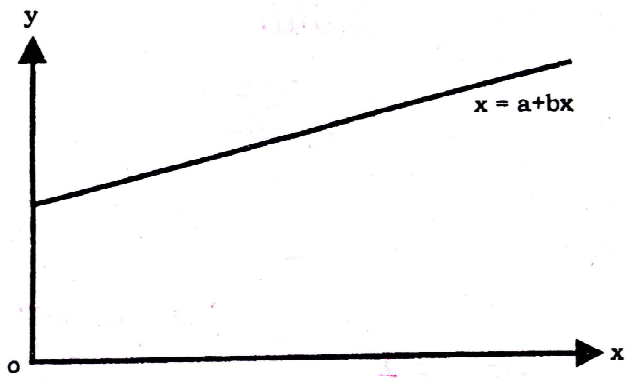
(a) Regression line of Y on X minimises total of the squares of the vertical deviations i.e. $\sum(y - y_c)^2$ is minimum. (See fig. 7).

(b) Regression line of X on Y minimises total of the squares of the horizontal deviations i.e. $\sum(x - x_c)^2$ is minimum (See fig. 8).



Regression line of y on x
$\Sigma(y-y_c)^2$ is minimum

(Fig. 7)



Regression lines of x on y
$\Sigma(x-x_c)^2$ is minimum

(Fig. 8)

1.4.3 Regression Equations :

Regression equations are the algebraic expressions of the regression lines. There are two regression lines, so there will be two regression equations.

(a) Regression equation of Y on X : Regression equation of Y on X describes the variation in the value of Y for the given changes in X. The regression equation of Y on X will be :

y = a + bx where X and Y are variables, and a and b are constants.

The 'a' constant is the y-axis intercept, i.e., the point where regression line touches the y-axis.

The constant 'b' shows the slope of the line.

(b) Regression equation of X on Y : Regression equation of X and Y describes the variation in the values of X for the given changes in Y. The regression equation of X on Y will be

X = a + by

Here, 'a' tells us how high above the X-axis the regression line is started and b = slope of the line.

Normal Equations : Two normal equations have to be solved for finding the values of constants a and b and in the regression equation Y on X i.e. Y = a + bx.

They are :

$\Sigma y = Na + b\Sigma x$

$\Sigma xy = a\Sigma x + b\Sigma x^2$

and for regression equation X on Y i.e. X = a + by, they are

$\Sigma x = Na + b\Sigma y$

$\Sigma xy = a\Sigma y + b\Sigma y^2$

The coefficient of regressions are found by the formula :

Regression coefficient of X on Y, $b_{xy} = r\ \frac{\sigma_x}{\sigma_y}$

Regression coefficient of Y on X, $b_{yx} = r\ \frac{\sigma_y}{\sigma_x}$

where σ_x = standard deviations of x series.

σ_y = standard deviations of y series.

r = coefficient of correlation of x and y series.

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{N\sigma_x\sigma_y} \times \frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{N\sigma_y^2} = \frac{\Sigma xy}{\Sigma y^2}$$

Similarly, $b_{yx} = \frac{\Sigma xy}{\Sigma x^2}$

Example 9.1 :

From the following data obtain the two regression equations :

x	:	5	8	7	6	4
y	:	3	4	5	2	1

Solution :

	x	y	x^2	y^2	xy
	5	3	25	9	15
	8	4	64	16	32
	7	5	49	25	35
	6	2	36	4	12
	4	1	16	1	4
Total (Σ)	30	15	190	55	98

Regression equation of Y on X : y = a + bx.

Now two normal equations are :

$\Sigma y = Na + b\Sigma x$ $\quad \therefore 15 = 5a + 30b$ ..... (i)

$\Sigma xy = a\Sigma x + b\Sigma x^2$ $\quad \therefore 98 = 30a + 190b$ ..... (ii)

Multiply (i) by 6, we get

$90 = 30a \pm 180b$ ..... (iii)

Substract (iii) from (ii)

Put b - .8 in (1) we get

15 = 5a + 30 (.8)

or 15 = 5a + 24

a = 1.8

By putting the values of a and b in equation, y = a + bx.

y = 1.8 + .8x.

Regression equation of X on Y : x = a + by.

Two normal equations are :

$\Sigma x = Na + b\Sigma y$ 30 = 5a + 15b ..... (i)

$\Sigma xy = a\Sigma y + b\Sigma y^2$ 98 = 15a + 55b ..... (ii)

Multiply (i) by 3 we get :

90 = 15a ± 45b

Again 98 = 15a + 55b ..... (ii)

and 90 = 15a ± 45b ..... (iii)

– – – –

Substracting (iii) from (ii) we get :

8 = 10b

b = .8

Put b = .8 in (i) we get

30 = 5a + 15 (.8)

or 30 = 5a + 12

∴ – 5a = 18or a = 3.6

Putting the value of a and b we get,

x = 3.6 + .8y.

Example 9.2 :

Two regression equations are

3x + 2y – 26 = 0 and

6x + y – 31 = 0. Find $\bar{x}, \bar{y}$ and r.

Also determine σ_y if $\sigma_x = 5$.

Solution :

Calculation of $\bar{x}, \bar{y}$

3x + 2y – 26 = 0 ..... (i)

6x + y – 31 = 0 ..... (ii)

Multiply (i) by 2 we get,

6x + 4y - 52 = 0

Substract (ii) from (iii) we get 3y - 21 = 0

or $y = \frac{21}{3} = 7 \ (\bar{y} = 7)$

substituting value of Y in (i) :

$3x + 2\times7 - 26 = 0$

$3x + 14 - 26 = 0$

$3x - 12 = 0 \quad \therefore x = 4 \quad (\bar{x} = 4)$

Calculation of 'r' : From eq (i) y/x : $2y = 26 - 3x$.

$y = 13 - 1.5 \times$ or $b_{yx} = .1.5$

x/y : $6x = 31 - Y$

or $x = 5.17 - .17y$

on $b_{xy} = .17$

$r = \sqrt{b_{xy}.b_{yx}} = \sqrt{-1.5 \times -.17} = \sqrt{255}$

r = -.5

SI of y : $b_{yx} = r \ .\sigma_y/\sigma_x$

$- 1.5 = -.5 \ \frac{\sigma_y}{5}$

or $-.5\sigma_y = -7.5 \quad \therefore \sigma_y = 15.$

1.4.4 Properties of Regression Coefficients :

1. The geometric mean between two regression coefficients is coefficient of correlation :

$$r = \frac{\sum xy}{\sqrt{\sum x^2.\sum y^2}} = b_{yx} \times b_{xy}$$

2. If one regression coefficient is greater than unity other regression coefficient must be less than unity.
3. Regression coefficients are independent of origin but not of scale.

Example 9.3 : If $b_{xy} = .8$ and $b_{yx} = .6$.

What would be the value of the coefficient of correlation.

Solution :

The value of coefficient of correlation is the geometric mean of the two regression coefficients. That is,

$r = \sqrt{.8 \times .6} = \sqrt{.48}$

Example 9.4 : Two regression equations are given as

x - 4y = -13 and 9y - x = 53 and $\sigma_x = 12$

We have to find (a) Mean of X and mean of Y.

(b) Coefficient of Correlation.

Solution : (a) Means of X and Y

$\bar{x}$ and $\bar{y}$ can be obtained by solving the given equations simultaneously for x and y.

The given equations are x - 4y = -13 ..... (i)

-x + 9y = 53 ..... (ii)

Adding (i) and (ii) we get 5y = 40 or y = 8.

Put y = 8 in (i), x - 32 = -13 or x = 19

(b) Coefficient of Correlations :

From the given equation, we do not know which regression equation is y on x and which regression equation is x on y. Hence, we have to take one equation as y on x and the other as x on y.

If $b_{yx} . b_{xy} \leq 1$, then our assumption is correct.

But if $b_{yx} . b_{xy} > 1$, then we change the assumption.

Let x - 4y = -13 is regression line of y on x.

and 9y - x = 53 is regression line of x on y.

From y on x equation.

x - 4y = -13

or 4y = x + 13

$$\therefore y = \frac{1x}{4} + \frac{13}{4}$$

Hence $b_{yx} = \frac{1}{4}$

From x on y equation.

9y - x = 53

x = 9y - 53

Hence $b_{xy} = 9$

Now $b_{yx} . b_{xy} = \frac{1}{4} \times 9 > 1$. This shows that our assumption is wrong.

1.4.5 Deviations Taken from Assumed Means :

When actual means of x and y variables are in decimals, the calculations can be simplified by taking deviations from the assumed means.

The two regression equations are

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

Where $r\frac{\sigma_x}{\sigma_y} = b_{xy} = \frac{\Sigma d_x d_y - \frac{\Sigma d_x d_y}{N}}{\Sigma d_y^2 - \frac{(\Sigma d_y)^2}{N}}$

$d_x = x - A$ and $d_y = y - A$.

Similarly, the regression equation of y on x is

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

$$r\frac{\sigma_y}{\sigma_x} = b_{yx} = \frac{\Sigma d_x d_y - \frac{\Sigma d_x d_y}{N}}{\Sigma d_x^2 - \frac{(\Sigma d_x)^2}{N}}$$

Example 9.5 : Given

x	:	6	2	10	4	8
y	:	9	11	5	8	7

x	**x-5**	**d_x^2**	**y**	**y-7**	**d_y^2**	**$d_x d_y$**
6	+1	1	9	2	4	2
2	-3	9	11	4	16	-12
10	+5	25	5	-2	4	-10
4	-1	1	8	1	1	-1
8	+3	9	7	0	0	0
$\Sigma x = 30$	$\Sigma d_x = 5$	$\Sigma d_x^2 = 45$	$\Sigma y = 40$	$\Sigma d_y = 5$	$\Sigma d_y^2 = 25$	$\Sigma d_x d_y = -21$

Regression equation of x on y : $x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$

$$\bar{x} = \frac{\Sigma x}{N} = \frac{30}{5} = 6, \bar{y} = \frac{\Sigma y}{N} = \frac{40}{5} = 8$$

$$r = \frac{\sigma_x}{\sigma_y} = \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N\Sigma d_y^2 - \left(\Sigma d_y\right)^2}$$

$$= \frac{5(-21) - (5)5)}{(5)(25) - (5)^2} = -\frac{105 - 25}{125 - 25} = \frac{130}{100} = -1.3$$

x-6 = .1.3 (y - 8)

x-6 = -1.3y + 10.4 or x = 16.4 - 1.3y

Regression Equation of y on x : $y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$

$$r = \frac{\sigma_y}{\sigma_x} = \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N\Sigma d_x^2 - \left(\Sigma d_x\right)^2}$$

$$= \frac{5(-21) - (5)5)}{(5)(45) - (5)^2} = -\frac{105 - 25}{225 - 25} = \frac{130}{200} = -.65$$

y - 8 = -0.65 (x - 6)

y - 8 = -0.65x + 3.90

y = -0.65x + 11.9 y = 11.9 - 0.65 x

SUGGESTED BOOKS

1.	Statistical Methods	:	S.P. Gupta
2.	Statistical Methods	:	C.B. Gupta
3.	Statistics	:	B.N. Gupta

Type Setting by Department of Distance Education
Punjabi University, Patiala