DESCRIPTIVE STATISTICS

BY

TANUJIT CHAKRABORTY

Indian Statistical Institute

Mail : tanujitisi@gmail.com

AN INTRODUCTION TO STATISTICS

STEPS AT A GLANCE:

Collection of Data –> Summarization of Data –> Analysis of Data –> Interpretation of Data towards a VALID DECISION.

WHAT IS THE MAIN PROBLEM IN STATISTICS?

Answer: Given a sample (a set of outcomes), we are to say (infer) about the population or the model. Statistics primarily deals with situations in which the occurrence of some event can't be predicted with certainty.

WHAT ARE THE MAJOR OBJECTIVES OF STATISTICS?

Answer: 1. To make inference about a population from an analysis of information contained in the sample data.

2. To make assessments of the extent of uncertainty involved in these inferences.

3. A third objective, no less important, is to design the process & the extent of sampling so that the observations from a basis for drawing valid & accurate inferences.

GIVE THE DEFINITION OF STATISTICS?

Answer: "Statistics" is a science of decision making on the basis of sample observations drawn from a population under uncertainty. That is, it is a mathematical discipline concerned with the collection of data, summarization of data, analysis of data & interpretation of data toward a valid decision.

Encyclopedia Americana: As a name of a field of study, Statistics refers to the science & arts of obtaining & analyzing quantitative data with a view to make sound inferences in the face of uncertainty.

Encyclopedia Britannica: As is commonly understood nowadays, Statistics is a mathematical discipline concerned with the study of masses of quantitative data of any kind.

WHAT IS THE MEANING OF THE TERM 'STATISTICS'?

Answer: As a singular noun it refers the science of collecting, analyzing & interpreting numerical data relating to an aggregate of individuals. As a plural noun it denotes the numerical & quantitative information, e.g., labor statistics, vital statistics.

IS STATISTCS A SCIENCE?

Answer: Any Science has for its objectives the formulation of laws for explaining phenomena in some part of the real world with a deterministic viewpoint. As Kendall explained, "Statistics is the branch of scientific method which deals with the data obtained by counting or measuring the properties of population of natural phenomena". Indeed, we can call Statistical Methodology as Scientific Method. It is noted that STATISTICS is sometimes called the study of variation, i.e., a population or group without any variation & uncertainty is no interest to Statistics. So, Statistics is the scientific methodology which deals with the collection, classification & tabulation of numerical facts as a basis for explanation, description & comparison of social phenomena.

Some Useful Definitions

Statistical Data: The numerical data or measurements obtained in case of an enquiry into a phenomenon, marked by uncertainty & variability, constitute Statistical Data. Uncertainty & variability are two major characteristics of Statistical Data. Not all quantitative data is statistical data. Example of statistical data – Suppose we study the 'Heights of students in a particular college'. Here we can't predict the height of an individual with certainty & there will be variation in heights of students. Counter Example: Multiplication table in a tabular form is a quantitative data, but since there is no uncertainty & variability involved in the data so it's not a Statistical Data.

Population: A set or group of observations relating to a phenomenon under statistical investigation is known as statistical population or simply population. However, the term 'population' implies an aggregate or collection of measurements on a given variable(s). Population is said to be finite or infinite according to whether the set contains a finite or infinite number of observations. Example – Measurements of heights in your college. Note that: 1. Characteristics of a population are called parameters. 2. A population contains finite or infinite no of observations on a given variable(s).

Sample: The set of data actually collected through a process of observation from selected items of any source is called a Sample. However, "Sample" is a subset of population or a true representation of population. Example – Measurements of heights of students of Statistics department in your college. Note that: 1. Characteristics of sample are called as Statistic. 2. A sample is taken in order to gather information about a population.

WRITE DOWN THE LIMITATIONS OF STATISTICS?

Answer: 1. Statistics deals with quantitative data only. 2. Statistical law holds good only for aggregate of items or average individuals. It may not be true for a particular individual or item. 3. Inadequate knowledge of data interpretation may lead to invalid decision. There are some saying: "There are three kinds of lies –lies, white lies & Statistics", "Figure won't lie, but liars figure", "Statistics is like a clay of which you can make a god or devil".

COMMENT ON THE FOLLOIWNGS:

1. **"In a study of ages & professions of deceased men, it was found that the profession with the lowest average age of death was 'student'. So it appears that student profession is very dangerous."**

Answer: It is obvious that every profession must have some basic education & it happens that the average age of every professional men must be higher than the age of the students. But it can happen that profession with lowest age of death was student. So the given statement is TRUE. But the conclusion made from the fact is incorrect. It can never be told that the student profession is dangerous. To conclude properly, we must have data for computing the proportions or percentages of deceased men in different professions. Therefore the conclusion made here is absurd.

2. **"A study of fighting in bars in which someone was killed found that 90% of the cases, the person who started fighting was the one who died".**

Answer: It is very obvious that the data is collected from the persons who has survived in fighting, not from the dies persons; as a matter of fact the survived persons will not accept the guilt of killing his opponent fighters. If the data is collected from eye witnesses, it is expected to say something in favor of the survived person as the opponent fighter has already been dead, from the sympathetic ground. Hence, the data collected may not be reliable at all. This is an examples of situation where inadequate information takes into bad decisions.

3. **"Blindly using any data happened to be available can lead to misleading information & bad decision."**

Answer: There are two kinds of people: Some of them believe that the inferences based on statistical data are very reliable & trusty. And others don't believe statistical results at all, they think it as a damned lies. But the fact is statistics is sometimes misused either deliberately or often due to lack of knowledge. Making conclusions based on inadequate information, deliberate manipulation & personal bias may lead to bad decision. Statistics are not to be blamed for all these. Statistics is like a clay of which you can make a god or a devil.

TYPES OF DATA

Statistical data may be classified in the following ways:

(i) Quantitative and Qualitative Data
(ii) Frequency data, Non−frequency data.
(iii) Nominal data, Ordinal data

By **Quantitative Statistical data** we mean a sequence or a set of numerical measurements made on some of the objects in a specified population. Therefore we may say that these types of data arise if we are observing, for each individual of a group, a character which can be measured in numbers. Such a character will be referred to as a quantitative character or a **variable.**

For example, the heights of 10 students of a college constitute quantitative data and the quantitative data are: 5'6", 5'5", 5'4", 5'8", 5'6", 5'7", 5'4", 5'5", 5'9" and 5'7". The character "Height of students" is a quantitative character or variable.

By **Qualitative Statistical data** we mean a set of observations in which each observation in the sample or population belongs to one of several mutually exclusive classes which are likely to be non−numerical. For this type of data, the character observed is not measurable in numerical terms. Such a character is called a qualitative character or an **attribute.**

The "color of a flower" can be classified as red, blue, white and others. The colors of ten flowers in a garden are recorded as: R, W, O, B, R, W, W, R, O and R. The data is qualitative data and the character "color of flower' is a qualitative character or an attribute.

Discrete and Continuous Variables:

When we study the data regarding quantitative characters, it is found that this may be of two types:

In the first type (Discrete Variable), the character may take only some isolated values, like the number of members in a family and the number of letters in a word, etc.

In the second type (Continuous Variable), the character can take any value within its range of variation. The height, weight, age of man are variables of this type. It is to be noted that in the second type, the actual measurements will present a discreteness, as for e.g.: when heights are given correct to the nearest 'cm'. But this discreteness is completely artificial, being due to the limitations of measuring instruments.

Variable of the first type are called discrete or discontinuous variables of the second type are called continuous.

Definition:

A variable which can take only some isolated value is called a **discrete variable.**

A variable which can take any value within its range of variation is called a **continuous variable.**

Non−frequency type data:

Time Series Data: When data are arranged according to the order of time, the data is known as time series data or historical data or chronological data. Here the values of one or more variables are given for different points or periods of time. Generally, in such a case, we are interested in the relationship between the time and the variable.

Example:

Time (in years)	Production of Rice in W.B. (in tons)
1990	10
1991	11
1992	11.5
1993	12
1994	10.5
1995	12
1996	13

Cross-sectional Data: It is a type of data which is collected by observing many subjects (such as individuals, firms, countries, or regions) at the same or approximately the same point in time, or without regard to differences in time.

Example: Here we study the changes in the value of the variable from a region to region.

States	Production of Rice (in tons)
Bihar	10
W.B.	16
Orissa	12
UP	14
MP	12
AP	11
Tamil Nadu	10

Frequency type data:

Consider the data on one or more variables for different individuals, may be even for different points of time, for different regions, but the identity of the individuals is not important and can be ignored. Now, we are interested in the characteristics of the group formed by the individuals rather than in those of the individual themselves. These types of data is called frequency type data, for here we are interested in knowing how frequently each of the different values of a variable occurs in a set of data.

Example: Marks distribution of First year students in a college:

Marks	0–30	30–40	40–50	50–60	60–70	70–80	above 80
No. of Students	20	30	40	50	40	20	30

Nominal data and Ordinal data:

Data may be divided into several categories. Categorical data have two primary type. Data having categories without a natural ordering are called **nominal data.** Examples are data on religion affiliation with the categories Catholic, Protestant, Hindu, Muslim and others. For nominal data, the order of listing the categories is irrelevant.

Many categorical data do have ordered categories, such data are called **ordinal data**. Examples are data on social class: upper, middle, lower; and on patient condition: good, fair and serious. Another example is data on political philosophy: liberal, moderate and conservative.

DIFFERNT TYPES OF SCALE

The theory of measurement consists of a set of separate or distinct theories, each concerning a distinct level of measurement. Here we will discuss four levels of measurement –nominal, ordinal, interval and ratio.

(a) The Nominal Scale

Definition: Measurement at its weakest level exists when numbers or other symbols are used simply to classify an object, person, or characteristic. When numbers or other symbols are used to identify the group to which various objects belong to, these numbers or symbols constitute a nominal scale.

Examples**:** The numbers on automobile license plates constitute a nominal scale. In India, a certain number or letter on license plate indicates the state in which the car owner resides, each subclass in the nominal scale consists of a group of entities: all owners residing in the same state.

In a nominal scale, the scaling operation is partitioning a given class into a set of mutually exclusive subclass. The members of any one subclass must be equivalent in the properly being scaled, that is, the only relation involved is that of equivalence.

(b) The Ordinal or Ranking Scale

Definition: It may happen that objects in one category of scale are not just different from the objects in other categories of that scale, but that they stand in some kind of relation to them. Typical relations among classes are: higher, more preferred, more difficult, etc. Such relations may be designated by carat (>), which means 'greater than'.

If the relation > holds for all pairs of classes so that a complete rank ordering of classes arise, we have an ordinal scale.

Examples: In prestige or social acceptability, all members of the upper middle class are higher than (>) all members of the lower middle class.

The fundamental difference between a nominal and an ordinal scale is that the ordinal scale incorporates not only the relation 'equivalence' (≡) but also the relation "greater than" (>). The scale is "unique upto a monotonic transformation" that is, it does not matter what numbers we give to a pair of classes, just as long as we give a higher number to the members of the class which is "greater" or "more preferred".

(c) The Interval Scale

Definition: When a scale has all characteristics of an ordinal scale and when in addition the distances between any two numbers on the scale are of known size. In such a case, measurement has been achieved in the sense of an interval scale. An interval scale is characterized by a common and constant unit of measurement which assigns a real number to all pairs of objects in the order set. In this sort of scale, the ratio of any twp intervals, is independent of unit of measurement and of the zero point. In an interval scale, the zero point and the unit of measurement are arbitrary.

Examples: We measure temperature on an interval scale. The unit of measurement and the zero point in measuring temperature are arbitrary –they differ for two scales: centigrade, Fahrenheit.

The interval scale is "unique up to a linear transformation", that is, the information yielded by the scale is not affected by the transformation $f(x) = ax + b$. In the temperature example,

$$F = \frac{9}{5}C + 32.$$

The interval scale is the first truly qualitative scale that we have discussed earlier.

(d) The Ratio Scale

Definition: When a scale has all the characteristics of an interval scale and in addition has a true zero point as its origin, it is called a ratio scale. In a ratio scale, the ratio of any two scale points is independent of the unit of measurement.

Examples: We measure mass or weight in a ratio scale. The scale of kg or pounds has a true zero point. The ratio between any two weights is independent of the unit of measurement. The operations of arithmetic are permissible on the numerical values assigned to the objects themselves.

COLLECTION OF DATA

Primary data and Secondary data:

The numerical facts or measurements obtained in the course of an enquiry into a phenomenon, marked by uncertainty, constitute statistical data. The statistical data may be already available or may have to be collected by an investigator or an agency.

The statistical data may be of two broad types: (i) Primary data, (ii) Secondary data.

Data is termed **primary** when the reference is to data collected for the first time by the investigator (or on his behalf). Data is termed **secondary** when the data is taken from records or data already available. The Meteorological Department regularly collects data on different aspects of the weather and climate such as amount of rainfall, humidity, maximum and minimum temperature of a certain place. These constitute **primary data.** To someone using them for a certain investigation the data will be **secondary data.**

Distinguish between Primary data and Secondary data:

(1) Primary data are those which are to be collected for the first time by the investigator (or on his behalf) and the therefore, it is of original in nature, whereas Secondary data are those which do not originate from the investigator (or from the field of enquiry) but which are obtained from someone else's records.

(2) Primary data may be used with greater confidence because the investigator will himself decide upon the coverage of the data, whereas secondary data is not so reliable. The secondary data may contain mistakes due to errors in transcription made when figures were copied.

There are two principal methods of data collection. Through a census, through a sample survey.

Census implies complete enumeration of each and every element of the source. Data obtained by taking relevant measurement or observation of each and every element of the source constitute **census data**. When only some selected elements of the source (selected according to some valid procedure) are taken and measurement or observations of these selected elements are recorded, the data is said to be collected through a sample enquiry and is said to be **sample data.**

The **advantages of Sample Survey method over the Census method** of enquiry are the following:

(1) Reduced Cost: The sample method is more economical.

(2) Greater scope: Complete enumeration is sometimes neither desirable nor feasible. In such cases only the sample method is to be adopted. Moreover, it is possible to collect more information in a sample enquiry than in a complete count.

(3) Greater speed: Data can be collected more quickly and summarized with a sample than with a complete count or census.

(4) Greater Accuracy: It is possible to engage better trained personnel for collection of data in the case of a sample enquiry than in a complete count. Processing of data is also much easier with sample data. All these factors lead to greater accuracy in data collected.

Methods of Collection of Data:

Statistical data are frequently obtained by a process in which the desired information is obtained from the source, either by having an enumerator visit to the informant, ask the necessary questions and enter the replies on a schedule, or by sending to the informant a list of questions (sometimes called a questionnaire) which he may answer at his convenience.

Questionnaire: The term 'questionnaire' means a list of certain systematically arranged questions relating to the subject of enquiry. It is necessary that questionnaire is designed with due care so that necessary data may be easily collected.

Schedule : In the schedule one finds a list of items, on which information will be collected, the exact forms of the questions to be put to the informants are not given and task of questioning, explaining the desired information is left to the investigator.

Framing of Questionnaire or Schedule:

Great care is to be taken in drafting a questionnaire or schedule, as this is the medium through which information is collected. Further it is also to be seen that the information collected is usable. Apart from care, expertise such as skill, wisdom, experience of the phenomenon under enquiry are needed in drafting a questionnaire. There are a few general points which should be borne in mind:

(i) The questions put should be clear, concise and unambiguous.

(ii) Delicate questions are to be put with greater care, often indirect questions should be put to get answers to some pertinent point. It is sometimes desirable to avoid very delicate questions.

(iii) The size of the questionnaire/schedule should be small. It saves time, both for the enumerator and the respondent. A large questionnaire is likely to exhaust the patience of the respondent.

(iv) There should be a natural, logical order in which questions are put.

(v) It should be noted that the information collected through questions should be such that it is usable.

Distinguish between Questionnaire and Schedule:

(a) A 'questionnaire' is a list of certain systematically arranged questions relating to the subject of enquiry, whereas a 'schedule' is a list of items on which information will be collected.

(b) The exact forms of the questions to be put to the informants are given, in 'questionnaire', whereas in schedule, the exact forms of the questions to be put to the informants are not given and task of questioning and explaining the desired information is left to the investigator.

(c) In questionnaire, the answers to questions are recorded by the informants himself, whereas in a schedule answers are recorded by the investigator.

Methods of Collection of Primary Data:

Primary data is collected through census or sample. There are several ways of collecting such data and these are

(i) The questionnaire method
(ii) The interviewer method
(iii) The method of direct observation

In the **questionnaire method**, each informant or respondent is provided with a questionnaire, usually sent by mail with return postage prepaid, and is asked to supply the information in the form of answers to the questions.

The merits of such a method are:

(1) It is much less time consuming and is economical.
(2) A much larger coverage can be made as, people in distant places can be reached without much difficulty.
(3) It is advantageous in a situation where the persons concerned move to faraway place.

The demerits are

(1) The method can be adopted only in case of educated people.
(2) The proportion of non−response is usually much larger. People do not have the time to spare nor are they willing to take the trouble of writing the answers themselves and of returning the questionnaire. Sometimes people also do not like to record information in their own handwriting and very often avoid answering delicate questions.

In the **interviewer method**, enumerators go from one informants to another and elicit the required information. This method is used in population census.

The merits of such a method are:

(1) Information so collected is more accurate, reliable and useful. The investigator can check and countercheck the information and get in the form in which he desires.
(2) The investigator can put alternative questions suited to the educational and cultural level of the persons concerned.
(3) In such cases, information can be collected by eliminating the bias and prejudices of the persons concerned.

The demerits are:

(1) Such a method can be adopted only when the enquiry is intensive and localized to a locality or a group. This cannot be used when the enquiry is extensive or is to be done in large areas.
(2) Such as enquiry is subjective in the sense that the intelligence, tact, skill as well as personal bias of the investigator are all reflected in the process.

In the method of **direct observation**, the enquirer or his assistant get the data directly from the field of enquiry without having to depend on the cooperation of informants. When data are needed on the height and weight of, say, 200 college students they will be approached individually and the height of each measured with a tape and the weight measured with weighing balance.

The merits of this method are:

(1) There is much lesser degree of subjectivity on the part of interviewer.
(2) Information so collected is more accurate, reliable.

The demerits are:

(1) It is expensive and time consuming.
(2) Thorough training of the enumerators is needed before they are sent to the field.

PRESENTATION OF DATA

The methods of collections of statistical data were described in the last part. The data so collected is known as raw data. The raw data which is, in general, huge and unwieldy, needs to be organized and presented in meaningful and readily comprehended form in order to facilitate further statistical analysis.

There are three broad ways of presenting data. These are: (1) Textual presentation, (2) Tabular presentation, (3) Graphic or diagrammatic presentation.

(1) Textual Presentation: In textual presentation, data is presented along with the text; that is, data may be incorporated in a paragraph of text.

Disadvantages: This is not a very effective and impressive device for statistical presentation, since it is necessary to read or at least scan, all of the paragraph before one can grasp the meaning of the entire set of figures, which takes up too much time. Most persons cannot easily comprehend the data presented in a paragraph of text, and it is especially difficult for the reader to single out individual figures. Data in this form are generally unarranged and unsystematic.

Advantage: There is the advantage, however, that the writer can direct attention to, and thus emphasize, certain figures and can also call attention to comparisons of importance.

(2) Tabular Presentation: In this presentation, data are arranged in a systematic way in rows and columns. Huge and unwieldy raw data can be neatly condensed in a table, by classifying data according to suitable groups or classes.

A table will have at least the four essential parts: title, stub, caption, body. There may also be present a footnote and prefatory note.

<u>Title:</u> A title should accompany every table and is customarily placed above the table. The title should be clearly warded and should state briefly what data are shown in the table.

When more than one table is included in a study, it is desirable to number the tables consecutively in order that each one may be identified by number rather than by title.

<u>Stub:</u> The extreme left part of the table; that is, the left hand column and its heading is called "stub", which is meant to describe the nature of the rows.

<u>Caption:</u> It is the heading of the other columns i.e., the upper part of the table which gives a description of the various columns is the caption of the table. The units of measurements for the data for each column are given in caption.

Body: The body is the principal part of the table, where all the relevant figures are exhibited.

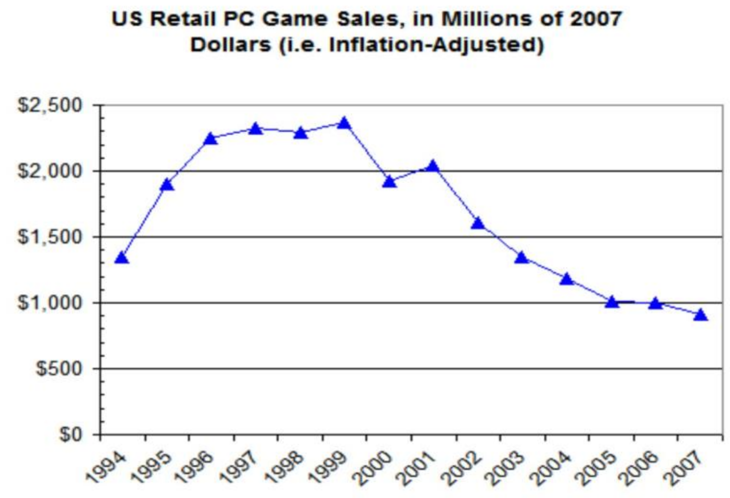
Footnotes and Prefatory Note: A Prefatory note, one or more footnotes and a source note may be appended to a table. The prefatory note is, placed just below the title and in small or less prominent type. The prefatory note provides an explanation concerning the entire table or a substantial part of it. Explanations concerning individual figures, or a column or row of figures, should be given in footnotes.

(3) Graphic or Diagrammatic representation of data: The important types of diagram which are used in statistical work are being described below.

(a) Line diagram:

When statistical data $\{(x_i, y_i), i = 1(1)n\}$ on two variables x and y are plotted in reference to x−axis and y−axis where both the axes are in arithmetic scale, the n pairs given n points on the graph, which are next joined by line segments. The resulting diagram is known as line diagram.

Line diagrams are frequently used for picturing time series data. Whenever time series data are represented by a line diagram, the time is shown in the x−axis and the other variable is placed on the y−axis. When considering the statistical data over a period of time, we are sometimes interested in the amount of change that has taken place, then the line diagrams are useful. Comparison of series differing not materially in the magnitude of individual items is possible with the line diagram.



Semi−logarithmic Chart:

A semi−logarithmic or ratio chart is a variant of line diagram where the vertical scale is logarithmic but the horizontal scale is of the arithmetic type. Comparison of series differing material in the magnitude of individual items is possible with the ratio chart. To compare the relative growth or decline of two or more series, one may use ratio chart.

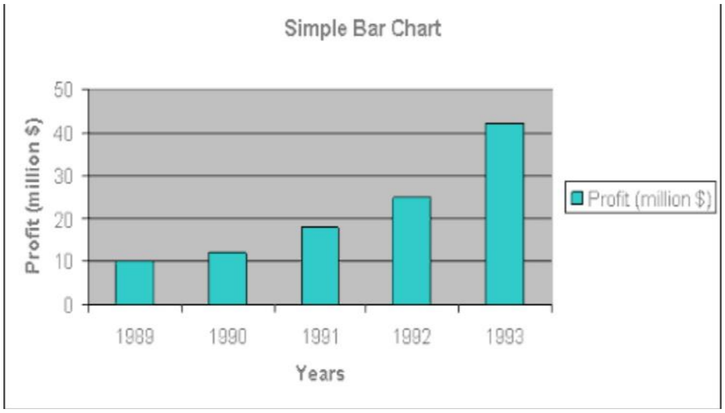
Logarithmic Scale: The construction of Logarithmic scale involves spacing the vertical scale values in proportion to the differences between their logarithms; that is, the vertical distance are proportional to the differences between the logarithms. An alternative approach to an understanding of the logarithmic scale does not involve logarithms is that equal distances measured along a logarithmic scale represent equal ratio.

(b) Bar diagrams:

Bar diagrams are the simplest and most used geometric forms for visual representation of data. Bar diagrams are of the following types:

(1) Simple bar diagram:

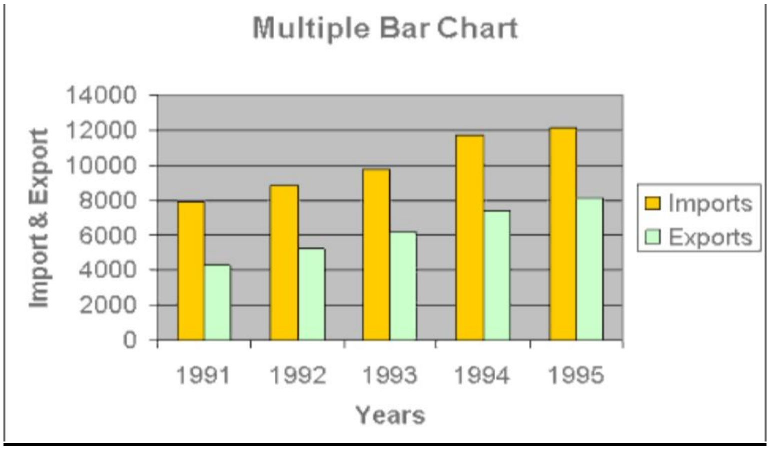
A Bar diagram which consists of a number of rectangles (usually called bars) is used for one−dimensional comparison. It is used to show absolute changes in magnitudes overtime (chronological) or space (geographical/regional). Changes in time or space, as the case may be, are shown along the x−axis with equally spaced magnitudes. Rectangles of equal width are drawn with lengths varying with the magnitude represented. While a line graph is not suitable for representation of data classified geographically or qualitatively, a bar diagram is suitable for representation of such data.



Vertical bars should also be used for data classified quantitatively. When making comparisons of data classified qualitatively or geographically, on the other hand, horizontal bars are generally used.

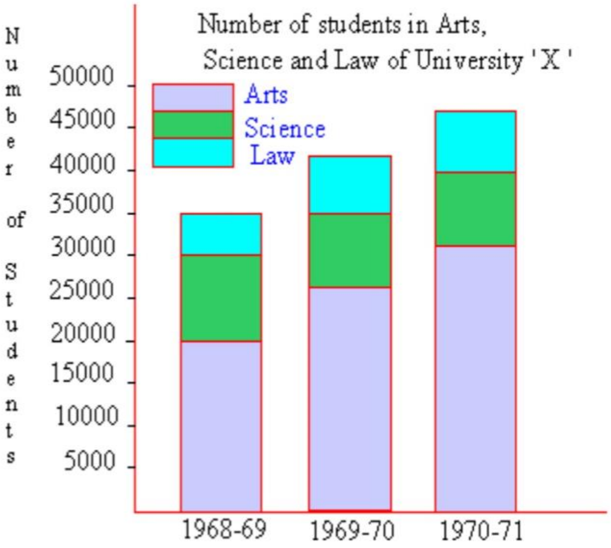
(2) Multiple Bar Diagram:

A multiple bar diagram is used for two or three dimensional comparison. For comparison of magnitudes of one variable in two or three aspects, or for comparison of magnitudes of two or three variables, a group of rectangles placed side by side is used. The bars are to be distinguished by shading or coloring to show the variables represented.



(3) Subdivided bar diagram:

The different components of a variable may be shown by subdivided bar diagram. Here as in the case of simple bar diagram, bars are drawn to represent the total magnitudes of the variable; one bar to represent each time period or geographical area. Then each bar is divided into several segments, each segment representing a component of the total. To distinguish between different components, different shading are used and explained in the body. The various components of the variable are to be represented in the same order in different bars to facilitate easy comparison.



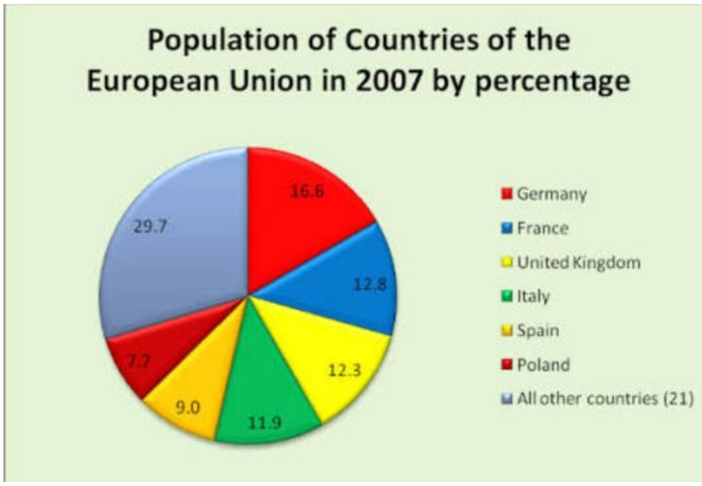
In the subdivided bar diagrams, the heights of the rectangles, drawn is proportional to the magnitudes of the variable.

Instead of considering absolute magnitudes of the variable, magnitudes of components may be indicated in terms of percentage of the total of the variable: the percentages are expected to give a better idea of the relative importance of each categories. For this purpose, a bar of suitable length is taken, its total area being regarded as 100 and then this area is divided in such a way that the area of each part represents the percentage for the corresponding category.

In order to draw a subdivided bar diagram, it is convenient to form beforehand a table of percentage.

(c) Pie diagram:

When an aggregate is divided into different components, we may be interested in the relative importance of the different components, rather than their absolute contribution to the aggregate. For representing breakdown of an aggregate into components a pie diagram is used. For pie diagram, one circle is used and the area enclosed by it being taken as 100. It is then divided into a number of sectors by drawing angles at the centre, the area of each sector representing the corresponding percentage. Since the full angle at the center is 360°, it is clear that for any particular category the angle (in degrees) should be 3.6 times the corresponding percentage.



FREQUENCY DISTRIBUTION

When observations on discrete or continuous variables are available on a single characteristic of a large number of members often it becomes necessary to condense the data as far as possible without losing any information of interest. If the data is non−frequency type, then the first step of condensation is to classify different values or is to divide the observed range of the variable into a suitable member of groups or classes, according to their increasing order in terms of magnitude and to record the number of observations corresponding to each distinct value or falling in each class.

Frequency: Number of times a variate value is repeated is called the frequency of the variate value. E.g., suppose there are 7 girl students who have secured 54 marks, 7 is the frequency of 54 marks.

The number of variate values belonging to a group or class is called the frequency of the class. If there are 12 people with monthly income Rs. 5000−7000, 12 is the frequency of the income group Rs. 5000−7000.

A table showing the distribution of the frequencies in the different values or classes is called a frequency table. A Frequency distribution shows how the total frequency is distributed over different variate values or different classes.

Frequency distributions:

(A) Discrete (Discontinuous) Variates:

We consider now how a frequency distribution table is to be constructed in the case of a discrete variable by taking a particular example.

Example: Four similar coins were tossed 20 times. The number of heads x in each of the 20 tosses were noted:

0, 1, 4, 2, 1, 3, 2

4, 0, 1, 2, 2, 3, 1

1, 2, 3, 2, 2, 3

Construct a frequency distribution table.

It should be noted that there is only five distinct values of the variable x, and they are 0, 1, 2, 3, 4.

Represent each value by a tally (/), for example, corresponding to a particular value 0, we put a tally (I) to the value $x = 0$;

Similarly we continue putting tallies for each value. We continue upto four tallies and the fifth tally us put cross wise (\) so that it becomes clear at once that the lot contains five tallies, i.e., there are five values. A gap is left after a lot of five tallies, before starting again to mark the tallies after each lot.

Table 1: Frequency distribution of number heads in tossing four coins

Values	Tally Marks	Frequency
0	//	2
1	/////	5
2	///// //	7
3	////	4
4	//	2
Total		20

The advantage of tally marks is that the single visit to the data is sufficient to construct or count the frequency.

(B)Continuous Variate:

We shall now consider construction of a frequency distribution table of a continuous variable.

Example: The heights of 50 students to the nearest centimeter are as given below:

151, 147, 145, 153, 156

144, 151, 157, 147, 150

151, 147, 155, 156, 151

149, 151, 153, 150, 152

151, 154, 155, 152, 154

152, 159, 153, 157, 152

157, 153, 151, 149, 147

158, 149, 147, 153, 152

154, 150, 152, 149, 151

152, 156, 155, 154, 150

Homework: Construct a frequency distribution table.

Class intervals and class limits: The interval defining a class is known as a class interval. For above example, 145−146, 147−148, …. are class intervals.

The end numbers describing a class interval are known as class limits; the smaller number is the lower class limit and the larger number is the upper class limit. The end numbers 145 and 146 of the class interval are the class limits; the smaller number 145 is lower class limit and the larger number 146 is the upper class limit.

Smoothening of a Grouped distribution:

If the data are being collected concerning heights of individuals, reported to the nearest centimeter, persons reported with heights 145 cm, would vary between 144.5 cm to 145.5 cm. For a frequency distribution of continuous variate, the class intervals do not constitute the continuous distribution, i.e., the upper limit of the previous class is not the lower limit of the following class, it has to be made continuous. The simple way to do this is to find the difference of the upper limit of the preceding class and lower limit of the following class. Subtract half of the difference from the lower limit of the following class and add the same to its upper limit. Continue this process for all classes.

Thus the smoothened frequency distribution will be:

Classes	Frequency
144.5−146.5	2
146.5−148.5	5
148.5−150.5	8
150.5−152.5	15
152.5−154.5	9
154.5−156.5	6
156.5−158.5	4
158.5−160.5	1
Total	50

Class boundaries and class width:

After making a frequency distribution continuous, i.e., in a smoothened frequency distribution, the end numbers of a class interval are called class boundaries; the smaller number is known as lower class boundary and the larger number is known as upper class boundary. The class boundaries of the 1st class: 144.5−146.5, are the numbers 144.5 and 146.5.

The difference between the upper and lower class boundaries is known as the width of the class. Here the width of the class: 145−146 or 144.5−146.5, is 146.5−144.5= 2cm and is the same for all classes.

General Rules for Construction of Frequency Distribution:

First, find the smallest and largest observations in the raw data supplied and find the range, i.e., the difference between the largest and the smallest observations.

Secondly, divide the range into a convenient number of class intervals having equal sizes. Sometimes it may be necessary to consider a slightly higher value than the exact range, so as to get a convenient number of class intervals of equal size. The number of class intervals taken should not be less than six or eight and greater than 15.

Another point to be borne in mind is that the midpoints coincide with actually observed data. However, whenever class boundaries are considered, it should be seen that no observation falls on the class boundaries. Sometimes the data is such that it is not possible to have all class intervals of equal size, in such case class intervals of unequal size, especially the class intervals at each end, may be conveniently taken.

Thirdly, find the numbers of observations falling in each class interval (or between corresponding class boundaries). This is best done by using tally marks.

Relative frequency and frequency density

Relative frequency of a class of a frequency distribution with n values is defined as

$$\frac{frequency\ of\ the\ class}{Number\ of\ values\ of\ the\ frequency\ distribution}$$

$$=\frac{frequency\ of\ the\ class}{n}$$

Frequency density of a class of a frequency distribution is defined as

$$\frac{frequency\ of\ the\ class}{width\ of\ the\ class}$$

General Frequency Distribution or Grouped data, with equal class width

Class boundaries	Class Marks	Frequency	Relative Frequency	Frequency Density
$\left(x_1-\frac{c}{2}\right)-\left(x_1+\frac{c}{2}\right)$	x_1	f_1	f_1/n	f_1/c
$\left(x_1+\frac{c}{2}\right)-\left(x_1+\frac{3c}{2}\right)$	x_1+c	f_2	f_2/n	f_2/c
$\vdots$ $\vdots$	$\vdots$ $\vdots$			
$\left(x_1+\frac{\overline{\iota-1}c}{2}\right)-\left(x_1+\frac{\overline{\iota+1}c}{2}\right)$	$\left(x_1+\frac{i}{2}c\right)$	f_i	f_i/n	f_i/c
$\vdots$ $\vdots$	$\vdots$ $\vdots$			
$\left(x_1+\frac{\overline{K-1}c}{2}\right)-\left(x_1+\frac{\overline{K+1}c}{2}\right)$	$\left(x_1+\frac{K}{2}c\right)$	f_K	f_K/n	f_K/c
Total		n	1	

Frequency Distribution of Grouped data with unequal class width

Class boundaries	Class marks	Frequency	Relative Frequency	Frequency Density
c_0--c_1	$\frac{c_0+c_1}{2}$	f_1	f_1/n	$f_1/(c_1-c_0)$
c_1--c_2	$\frac{c_1+c_2}{2}$	f_2	f_2/n	$f_2/(c_2-c_1)$
c_2--c_3	$\frac{c_2+c_3}{2}$	f_3	f_3/n	$f_3/(c_3-c_2)$
$\vdots$ $\vdots$ c_i--c_{i+1} $\vdots$ $\vdots$	$\vdots$ $\vdots$ $\frac{c_i+c_{i+1}}{2}$ $\vdots$ $\vdots$	$\vdots$ $\vdots$ f_i $\vdots$ $\vdots$	$\vdots$ $\vdots$ f_i/n	$\vdots$ $\vdots$ $f_i/(c_{i+1}-c_i)$
$c_{K-1}--c_K$	$\frac{c_{K-1}+c_K}{2}$	f_K	f_K/n	$f_K/(c_K-c_{K-1})$
Total		n	1	

Graphical representation of a Frequency Distribution:

It is convenient to represent the frequency distribution by means of a diagram which conveys the salient features of the data at a glance. It facilitates the comparison of two frequency distributions.

The following types of graphs are used in representing frequency distributions:

Frequency – Polygon:

For an ungrouped frequency distribution, measuring variate−value along x−axis and frequency along y−axis, and plotting points with abscissa as the variate values and the ordinates as the corresponding frequencies; joining the plotted points by straight line, one to the next, the diagram so obtained is called a frequency Polygon.

For a grouped frequency distribution, we erect at the abscissa corresponding to the centre of each class−interval an ordinate equal to the frequency per unit interval in that interval. The ends of these ordinate are joint by straight line, one to the next.

Table 1: Frequency distribution showing the number of boys on the registers of primary schools in a State

Age (in years)	No. of boys (in thousands)	Frequency per unit interval (Frequency density)
2−5	150	$150/(5-2) = 50$
5−9	2066	$2066/(9-5) = 518.5$
9−12	1497	$1497/(12-9) = 499$
12−13	477	477
13−14	496	496
14−15	143	143
15−16	162	162

Histogram:

For a given grouped frequency distribution, we first mark off along the x−axis all the class−interval on a suitable scale. With the class−intervals a bases we draw rectangles with areas proportional to the frequencies of the class intervals. For equal class intervals, the heights of the rectangles will be proportional to the frequencies, while for unequal class intervals, the heights will be equal (or proportional) to the frequency densities of the classes.

Homework: Draw Frequency Polygon & Histogram from the Data given in Table 1.

FREQUENCY CURVES:

If the class intervals be made smaller, and at the same time the number of observations increased so that the class frequencies may remain sizeable (or finite), the frequency polygon and the histogram will approach more and more closely to a smooth curve. Such an ideal limit to the polygon or the histogram is called a frequency curve. The frequency curve can be obtained by drawing a smooth free hand curve through the vertices of the frequency polygon, or through the midpoints of the top of the rectangles of the histogram.

In the frequency curve the area between the ordinates whatever is proportional to the number of observations falling between the corresponding values of the variable. Thus, the number of observations falling between the values of the variable x_1 and x_2 in the above figure will be proportional to the area of the shaded strip.

SOME COMMON TYPES OF FREQUENCY CURVE:

(1) The symmetrical or Bell Shaped

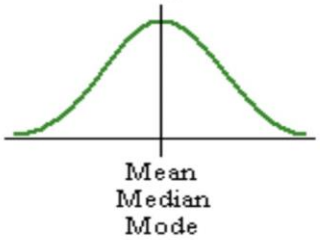
In this type the class−frequencies decrease to zero symmetrically on either side of a central maximum. (Graph 1)

(2) The moderately asymmetrical (skew)

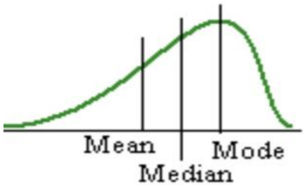
In this case the class frequencies decrease with markedly greater rapidity on one side of the maximum than on the other.

Asymmetrical curves are also said to be “skew”. If the longer tail of a curve lies to the right, the curve is called (positively skewed) skewed to the right or (negatively skewed) skewed to the left if it lies to the left. (Graph 2 & 3)

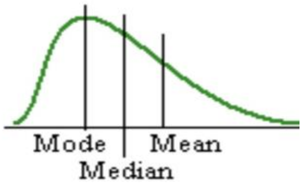
Graph 1



Graph 2



Graph 3



(3) The extremely asymmetrical or J−shaped

In this type the class−frequencies run up to a maximum at one end of the range.

Consider the frequency distribution of deaths from scarlet fever for 5 years intervals, the maximum number of deaths occur at the beginning of life i.e., in the age group 0−5, the distribution is J−shaped.

(4) The U−Shaped

This type exhibits a maximum frequency at the ends of the range and a minimum towards the centre.

CUMULATIVE FREQUENCY AND CUMULATIVE FREQUENCY CURVE

The number of observations which are less than or equal to a specified value x is called the cumulative frequency of x of "less than" type. The number of observations which are greater than or equal to a specified value x is called the cumulative frequency of x of "greater than" type.

Discrete variate:

Let a discrete variate x takes the value x_i with frequency $f(x_i), i = 1, 2, 3, \ldots.$

Then $F(x) = \sum_{x_i \leq x} f(x_i)$, the total frequency of less than or equal to x, is known as cumulative frequency of x of less than type.

Clearly,

$$F(x) = \begin{cases} 0 & if\ x < x_1 \\ f_1 & if\ x_1 \leq x < x_2 \\ f_1 + f_2 & if\ x_2 \leq x < x_3 \\ f_1 + f_2 + f_3 & if\ x_3 \leq x < x_4 \\ & \vdots \\ & \vdots \end{cases}$$

F(x) is a discontinuous function at each $x_i, i = 1, 2, 3, \ldots.$ The function F(x) is called the Step function. The graph or curve of F(x) is known as step diagram.

Continuous variate: (Grouped frequency distribution)

Consider the number of all observations which are less than or equal the upper class boundary of a given class interval: this number is the sum of the frequencies up to and including that class to which the upper boundary corresponds. This sum is known as the cumulative frequency up to and including that class.

Table: Cumulative frequency (less than) table of heights of 50 students:

Class interval	Frequency	Cumulative frequency	
		Less than	greater than
145–146	2	2	50
147–148	5	7	48
149–150	8	15	43
151–152	15	30	35
153–154	9	39	20
155–156	6	45	11
157–158	4	49	5
159–160	1	50	1
Total			

The cumulative frequency distribution is represented by joining the points obtained by plotting the cumulative frequencies along the vertical axis and the corresponding upper class boundaries along the x−axis. The corresponding polygon obtained by joining the points by straight lines, is known as cumulative frequency polygon or **ogive** (less than type).

Similarly we can construct another cumulative frequency distribution (more than type) by considering the sum of frequencies greater than the lower class boundaries of the classless. The graph obtained by joining the points obtained by plotting the cumulative frequencies (more than) along the vertical axis and the corresponding lower class boundaries along the x−axis is known as **cumulative frequency polygon or ogive** (“greater than” type).

STEM & LEAF PLOTS & BOX PLOTS ①

[CHAPTER-1 (Continuation)]

Stem & Leaf Plots :- This a method of displaying a set of data and it is a display that organises data to show it's shape and distribution.

Suppose that the data are represented by $x_1, x_2, \ldots, x_n$ and that each number x_i consist of at least two digits. To construct a stem and leaf plot, we divide each no x_i into two parts: A stem, consisting of one or more of the leading digits & leaf, consisting of the remaining digits.

Example:- Math test scores out of 50 points:

35, 36, 38, 40, 42, 44, 45, 45, 47, 48, 49, 50, 50, 50

Writing the data in numerical order may help to organize the data, but is not a required step. Ordering can be done later.

Seperate each number into a stem and leaf since these are two digit numbers, the ten digit is the stem and unit digits are the leaf.

The number 38 would be represented as,

Stem	Leaf
3	8

Group the numbers with the same stems. List the stems in numerical order. Title the graph.

Math Test Score (out of 50 mark

Stem	Leaf
3	5 6 8
4	0 2 2 4 5 5 7 8 9
5	0 0 0

Prepare an appropriate legend (key) for the graph.

Legend : 3|6 means 36.

Remark:- 1. (a) A stem & leaf plot shows the shape and distn of data. It can be clearly seen in the diagram above that the data clusters around the rows with stem of 4.

(b) the stem & leaf display, like the histogram, summarises the shape of a batch of data.

2. We prefer the stem & leaf display to the histogram because it remains the most significant digits of the data. This feature enable us:

(a) To see patterns in the data.
(b) To see the distribution of data values within an interval.
(c) To go more easily from a value in the display to the datumn that produced it.

The stem & leaf display & the distogram provide a visual impression about a set of data whereas the sample average, standard deviation provide quantitative information about specific features of the data.

Comparison between Stem and leaf display and Histogram:—

In explosing the analyst to the different features of data, the stem and leaf display has much in common with its close relative, the histogram.

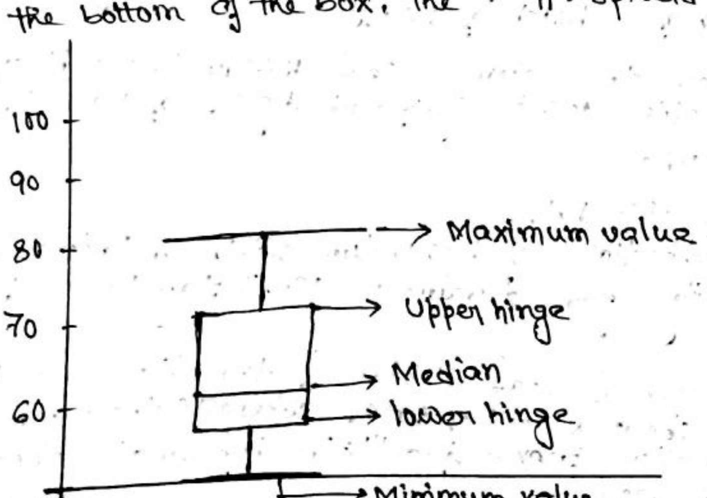
(i) When we work by hand, it is easier to construct stem-and-leaf display compare to draw a histogram.

(ii) By using the digits of the data values themselves, we construct the stem-and-leaf display whereas by constructing a groupted frequency distribution, a histogram is drawn by enclosing the area between the two class boundaries.

(iii) In stem-and-leaf display, it is easily possible to go from the display to the datum that produces it, but in constructing histogram, we form a frequency distribution and the original data is not obtainable from the histogram.

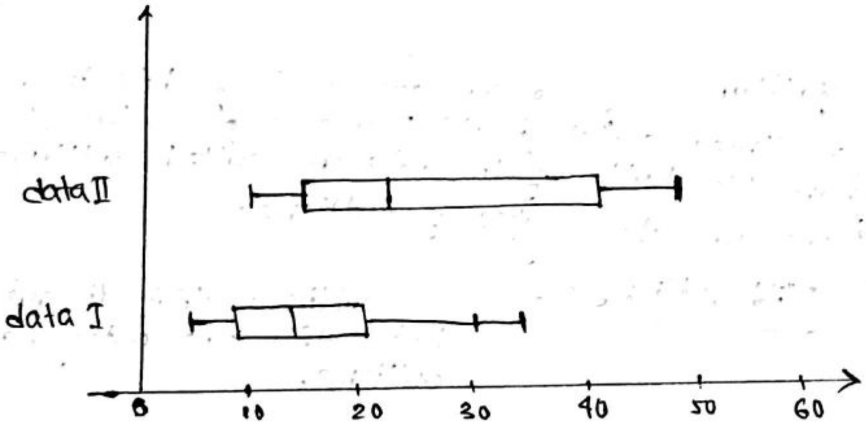
BOX-PLOT :— The box-plot is a graphical display, that simultaneously displays, several important features of the data such as. location or central tendency, spread or dispersion, departure from symmetry and identification of observations that lie ususually far from the bulk of the data (these observations are often called OUTLIERS). A box-plot displays three quartiles, Minimum & Maximum of the data on a rectangular box. The box stretches from the lower hinge (defined as the 25th percentile) to the upper hinge (defined as the 75th percentile) & therefore contains the middle half of the scores (data) in the distribution.

The median is shown as a line across the box, therefore 4th of the distribution is between this line and the top of the 4 box & 1/4 th of the distribution is between this line & the bottom of the box. The " H-spread" is defined as the difference between the hinges & a step is defined as 1·5 times the H-spread. A line at either end extends to the extreme values. the's lines are usually called whiskers. This compact visual display is useful for comparing several batches. By box-plot we can compare the batches location, spread & perhaps skewness & kurtosis.



The location of the data is summarized by the median, the crossbar in the interior box. The length of the box shows the spread, using H-spread. If the median is closer to the lower hinge than to the upper hinge, indicating that the data is positively skewed. Smaller the H-spread compare to the length {max - min.}, larger the kurtosis of the data.

Comparing data Using Box-Plots:- Box-plot display is especially useful for comparison of several data sets. By drawing a box-plot for each batch and arranging them in parallel, we can see similarities and differences among the data sets w.r.t. each of the features: Location, spread, skewness, kurtosis.



i) Clearly, the median of the data II is greater than the median of the data I, from the display. The average of data II is larger than that of the data I.

ii) The distance between the upper hinge and the lower hinge is a measure of spread. From display, the H-spread of data II is larger than the H-spread of data-I. Hence the spread or dispersion of data II is more.

iii) If a data is symmetric, then the upper-hinge - median = median - lower hinge. For the display of data I, they are symmetrically distributed. For the display of data II, the distance between upper hinge and median is greater than the distance between lower hinge and median, the data II is positively skewed.

iv) If H-spread is small compare to {maximum - minimum}, the distribution has high kurtosis. Clearly, the H-spread is small compare to {max - min} in the box-plot of data-I in comparison with data II. Hence, the kurtosis of data I is high compared to that of data-II.

Outliers:- A common problem that every statistician has to face in the course of work is to decide whether one or more of the observations available to him come from a distribution different from the distribution yielding the other observations. The enquirer scrutinises the data and gets the impression that some of the observations are too high or too low to be compatible with the assumption that they have been obtained from the same distribution. These observations are called outliers. Before studying the features of the distn. we should eliminate the outliers from the data to get proper measures of the features. In the absence of an outlier, we can use median or trimmed mean as a measure of location.

Definition:- If we look for a set of data in a distribution, we sometimes see that there are a few values lie on usually far from the bulk of the data are called outliers.

There are two types of outliers:

i) Mild Outliers :- The outliers that lie outside the innerfence but within the outerfence are called mild outliers.

ii) Extreme Outliers:- The outliers that lie outside the outerfence are called extreme outliers.

* ———————— *

Measure of Central Tendency

C.U

⇨ Define Central Tendency? or, What do you mean by Centoral Tendency of a frequency distribution?

Ans:→ A set of observations shows a tendency or motive to have a value (generally centrally located) by which they may be replaced. This character is termed as central Tendency.

For any frequency distribution we find a tendency of the variate values to cluster around a central value; in other words, most of the values lies in a small interval about a central value. this characteristic is called the central tendency of a frequency distribution. In relation to a frequency distribution, an average is also termed as a measure of location, because it helps to locate the position of the distribution on the axis of the variable.

⇨ Measures of Central Tendency.

Ans:→ Central tendency is measured by —

i) Mean,
ii) Median,
iii) Mode,
iv) Quartile,
v) Decile,
etc......

⇨ Arithmatic Mean.

For Non Frequency or raw data.

The arithmatic mean of a variable is derived by dividing the sum of its values by the no. of values. If x denotes the variable under consideration and its values namely $x_1, x_2, \ldots, x_n$ are given, then the arithmatic mean of x, denoted by $\bar{x}$, is given by

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

Scanned by CamScanner

Note:→ The computation of the arithmatic mean, in some cases, is simplified by subtracting a suitable factor c, say, from each observation.

Suppose $y_i = x_i - c$, for each i,

or., $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n}(x_i - c)$, where n denotes number of given values.

or, $\sum_{i=1}^{n} y_i/n = \sum_{i=1}^{n} x_i/n - nc/n$

or, $\bar{y} = \bar{x} - c$

Then, $\bar{x} = \bar{y} + c$.

For frequency Data.

For Discrete variable

If the values of a discrete variable are exhibited alongwith their corresponding frequencies, then the mean can be obtained in the following way :

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_n f_n}{f_1 + f_2 + \cdots + f_n}$$

$$= \sum_{i=1}^{n} x_i f_i / N \quad , \text{where } N = \sum_{i=1}^{n} f_i \text{ the total frequency.}$$

where $x_1, x_2, \ldots, x_n$ denote the distinct values of the variable x and $f_1, f_2, \ldots, f_n$ indicate their respective frequencies.

Note:→ Data ; $x_1, x_2, \ldots, x_n$

New data: $\bar{x}, \bar{x}, \ldots, \bar{x}$, where $\bar{x} = \frac{1}{n}\sum x_i$

Error : $x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$.

$$\begin{aligned} \text{Total Errors} &= \sum(x_i - \bar{x}) \\ &= \sum x_i - \sum \bar{x} \\ &= \sum x_i - n\bar{x} \\ &= \sum x_i - \sum x_i \\ &= 0. \end{aligned}$$

So, if we replace each observation by its mean, we arenot doing any errors, i.e. if observations are replaced by its mean, the observation remained unaffected.

For continuous Variable.

Again, for a continuous variable, the data are summarised in a frequency table showing the various class intervals and their corresponding class frequencies. In this case, the class-mark of a class-interval is supposed to represent the interval and on the basis of this assumption, an approximate value of the mean may be obtained. Hence the mean ($\bar{x}$) is expressed in the form

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} x_i f_i}{N}$$

In the case of equal width of the class intervals, calculation of the mean may be facilitated through a change of origin (or base) and scale. We are to subtract c from each class-mark and then divide the resultant by d, where c is the chosen origin, usually a class-mark near the middle of the range and d, the scale, is the common width. If y_i be the new value corresponding to x_i, then

$$y_i = \frac{x_i - c}{d}$$

or, $x_i = c + d y_i$, for each i

or, $x_i f_i = c f_i + d y_i f_i$, for each i

or, $\sum_i x_i f_i = c \sum_i f_i + d \sum_i y_i f_i$

or, $\frac{1}{n} \sum_i x_i f_i = c + \frac{d}{n} \sum_i y_i f_i$, where $n = \sum_i f_i$

or, $\bar{x} = c + d\bar{y}$.

Calculation of Mean :→

Class Boundaries	Frequency f_i	Class mark x_i	$y_i = \frac{x_i - 55.5}{10}$	$y_i f_i$
30.5–40.5	6	35.5	−2	−12
40.5–50.5	14	45.5	−1	−14
50.5–60.5	20	55.5	0	0
60.5–70.5	7	65.5	1	7
70.5–80.5	3	75.5	2	6
Total =	N=50	—	—	−13

Here, $\bar{y} = \sum_i y_i f_i / N$, where $N = \sum f_i$

$$= \frac{-13}{50} = -0.26.$$

Since $x_i = 55.5 + 10 y_i$,

$$\bar{x} = 55.5 + 10\bar{y}$$
$$= 55.5 + 10(-0.26)$$
$$= 52.9.$$

Some important properties of AM :→

(a) If the observed values of a variable are all equal, then their mean will be the common value.

Suppose we are given n values $x_1, x_2, \ldots, x_n$ of a variable x, where $x_i = c$, for each i.

Then $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} c = nc$.

Hence $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n} \times nc = c$.

(b) The sum of the deviations of the values of a variable from its mean is zero.

Case I: Suppose a variable x assuming n values $x_1, x_2, \ldots, x_n$ has mean $\bar{x}$, where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\Rightarrow \sum_{i=1}^{n} x_i = n\bar{x}$$

Now, $\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x}$

$$= \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0$$

Case II: For discrete frequency distribution, we get —

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i f_i \text{, where } N = \sum_{i=1}^{n} f_i$$

or, $\sum_{i=1}^{n} x_i f_i = N\bar{x}$.

Now, $\sum_{i=1}^{n} f_i(x_i - \bar{x}) = \sum_{i=1}^{n} f_i x_i - \bar{x}\sum_{i=1}^{n} f_i$

$$= N\bar{x} - N\bar{x}$$
$$= 0.$$

(c) Suppose x is a linear function of y in the form $x = a+by$, then the arithmatic means of x and y are related as $\bar{x} = a+b\bar{y}$.

Here, $x = a+by$ or $x_i = a+by_i$, for each i

$$\text{or, } \sum_{i=1}^{n} x_i = \sum_{i=1}^{n}(a+by_i)$$, where n denotes the number of given values

$$\text{or, } \sum_{i=1}^{n} x_i = na + b\sum_{i=1}^{n} y_i$$

$$\text{or, } \frac{1}{n}\sum_{i=1}^{n} x_i = a + \frac{1}{n}\sum_{i=1}^{n} y_i \cdot b$$

$$\text{or, } \bar{x} = a+b\bar{y}.$$

~~This result shows that AM is affected by change of both origin~~

(d) If there are two groups of values of variable x, one containing n_1 values with mean $\bar{x}_1$ and the other containing n_2 values with mean $\bar{x}_2$, then the mean of the combined data is given by

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1+n_2}.$$

Let the values in the first group be $x_{11}, x_{12}, \ldots, x_{1n_1}$, while those in the second group be $x_{21}, x_{22}, \ldots, x_{2n_2}$.

$$\text{Clearly, } \sum_{i=1}^{n_1} x_{1i} = n_1\bar{x}_1 \text{ and } \sum_{j=1}^{n_2} x_{2j} = n_2\bar{x}_2.$$

Hence the sum of all the values in the two groups, taken together, is given by

$$\sum_{i=1}^{n_1} x_{1i} + \sum_{j=1}^{n_2} x_{2i} = n_1\bar{x}_1 + n_2\bar{x}_2.$$

Then the mean of the combined data is

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1+n_2},$$ $\bar{x}$ is called Grand Mean or Combined Mean or composite mean.

Remarks:→

i) If there are k groups of values of a variable x such that these groups contain $n_1, n_2, \ldots, n_k$ values and have means $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$ respectively, then the grand mean of x is

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i\bar{x}_i}{\sum_{i=1}^{k} n_i}$$

ii) In particular, if $n_1 = n_2 = \cdots = n_k$, then $\bar{x} = \frac{\bar{x}_1 + \bar{x}_2 + \cdots + \bar{x}_k}{k}$

i.e. the mean of the combined data is equal to the mean of the means of the individual groups when there is an equal number of values in each group.

(e) If a variable u is related to two variables x and y as $u = ax + by$, then its mean is related to the means of x and y in the similar way, i.e. $\bar{u} = a\bar{x} + b\bar{y}$.

Let n pairs of values $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ of x and y be given. Then $u_i = ax_i + by_i$, for each i

$$\text{or, } \sum_{i=1}^{n} u_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} y_i$$

$$\text{or, } \frac{1}{n}\sum_{i=1}^{n} u_i = a.\frac{1}{n}\sum_{i=1}^{n} x_i + b.\frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\text{or, } \bar{u} = a\bar{x} + b\bar{y}.$$

Cor. If $a = 1, b = 1$, then $u = x + y$ and $\bar{u} = \bar{x} + \bar{y}$.

Arithmatic Mean :->

Merits :->

(a) It is rigidly defined.
(b) It is easy to understand and calculate.
(c) It is based on all observations.
(d) It is least affected by sampling fluctuations.
(e) It is amenable to algebric operations.

Demerits :->

(a) It can not be determined by inspection.
(b) It can not be calculated if variable under study is qualitative in nature.
(c) It is difficult to calculate if one or more observations are missing.
(d) It is affected by extreme observations.
(e) It can not be calculated if the two terminal points are open.

⇨ Median :→

C.U

⇨ What is Median?

Ans:→ The median of a variable is defined as the middlemost value when its values are arranged in ascending or decending order of magnitude. In other words, the median divides the whole set of values in two parts such that half of the observations are less than or equal to it and half are more than or equal to it.

n → odd median($\tilde{x}$) = $\left(\frac{n+1}{2}\right)$th value in the arrangement.

n → even median($\tilde{x}$) = $\frac{n}{2}$th or $\left(\frac{n}{2}+1\right)$th or the AM of $\frac{n}{2}$th and $\left(\frac{n}{2}+1\right)$th values in the ordered arrangements.

⇨ How is median computed from discrete and continuous frequency distributions?

C:U

Ans:→ For Discrete Frequency Distribution.

In connection with the frequency distribution of a discrete variable, the cumulative frequencies indicate an arrangement of the different values in an ascending or decending order of magnitude, depending on their type (i.e. less than or more than). Let us consider a variable x which assumes five distinct values $x_1, x_2, \dots, x_5$ with $F_1, F_2, \dots, F_5$ as their corresponding less than type cumulative frequencies; then it means that the first F_1 values are all equal to x_1, (F_1+1)th value or F_2th value are all equal to x_2 and so on. Here F_5 indicates the total number of obsn.s and on the basis of its even or odd value, the median of x can be obtained.

Example:→ Discrete data :→

x	f	cumulative frequency ($\le$)
1	1	1
2	3	4
3	7	11
4	4	15
5	2	17
	$n=17$	

$\therefore \tilde{x} = \frac{17+1}{2} = 9$th observation.

For Continuous Frequency Distribution.

In relation to the frequency distribution of a continuous variable, the median is regarded as the value for which the cumulative frequency is $\frac{n}{2}$. On observing the less than type, say, cumulative frequencies, we can obtain the class interval that contains the median. In fact, the cumulative frequency for this interval is just more than or equal to $\frac{n}{2}$. The value of the median can be approximately obtained by the following procedure.

Let us denote the lower and upper class-boundaries of the class containing the median by x_l and x_u and the corresponding cumulatible frequencies by n_l and n_u, respectively. If we assume that cumulative frequency changes from n_l to n_u between x_l to x_u at a constant rate, i.e. if we assume that cumulative frequency is a linear function of x between x_l to x_u, then the median, which is the value with cumulative frequency $\frac{n}{2}$, will satisfy the relation —

$$\frac{M_i - x_l}{x_u - x_l} = \frac{n/2 - n_l}{n_u - n_l}$$

This gives, $M_i = x_l + \frac{n/2 - n_l}{f_0} \times c$,

where c and f_0 are the width and the frequency of the class-interval containing the median, M_i.

The same value may also be obtained geometrically, from the ogive of the frequency distribution. The median will be given by the abscissa of the point on the ogive for which the ordinate is $n/2$.

Example: → See in the next page :-

	x	f		Cumulative frequency ($\le$)	
	30·5-40·5	6		6	
	40·5-50·5	14		$\boxed{20}$	$\rightarrow n_l$
median class ←	$\boxed{50·5-60·5}$	$\boxed{20}$	$\rightarrow f_0$	40	
	60·5-70·5	7		47	
	70·5-80·5	3		50	

$C = 10$ $\quad n = 50$

x_l = lower boundary of the median class = 50·5 , $n_l = 20$

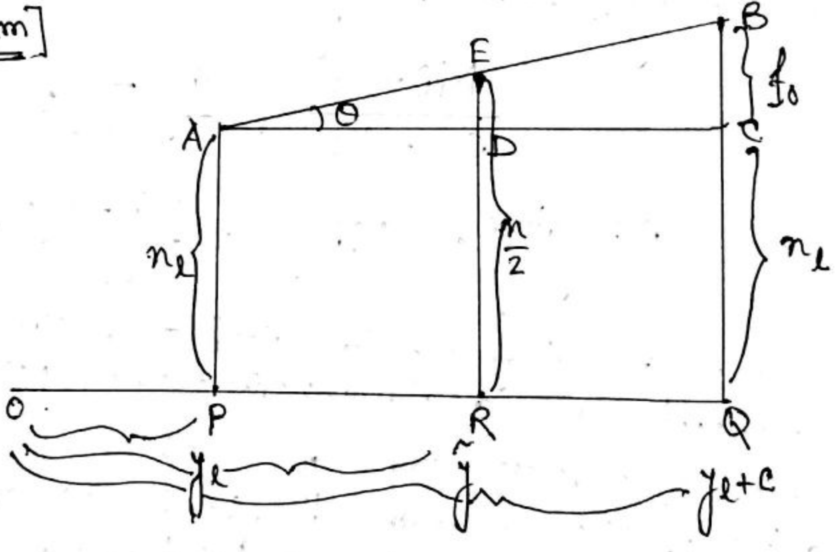
$$M_i = \tilde{x} = x_l + \frac{\frac{n}{2} - n_l}{f_0} \times C$$

$$= 50{\cdot}5 + \frac{\frac{50}{2} - 20}{20} \times 10$$

$$= 53.$$

☆ **Result** :→ Proof that – $\tilde{y} = y_l + \frac{\frac{n}{2} - n_l}{f_0} \times c$

Proof :→ [Not for exam]



$$\tan\theta = \frac{ED}{AD} = \frac{BC}{AC}$$

$$\Rightarrow \frac{ER - DR}{OR - OP} = \frac{BQ - CQ}{OQ - OP}$$

$$\Rightarrow \frac{\frac{n}{2} - n_l}{\tilde{y} - y_l} = \frac{\cancel{n_l} + f_0 - \cancel{n_l}}{\cancel{y_l} + c - \cancel{y_l}}$$

$$\Rightarrow \boxed{\tilde{y} = y_l + \frac{\frac{n}{2} - n_l}{f_0} \times c}$$

Scanned by CamScanner

An important property of median :—

If two variables x and y be linearly related in the form $y = a + bx$, then $Me(y) = a + b.Me(x)$.

Suppose we are given n values $x_1, x_2, \ldots, x_n$ of variable x, such that

$$x_1 \le x_2 \le x_3 \le \cdots \le x_n . \text{ —— (i)}$$

Let $y = y_i$ when $x = x_i$. then we get,

$$\left.\begin{array}{ll} y_1 \le y_2 \le \cdots \le y_n & \text{when } b > 0 \\ y_1 \ge y_2 \ge \cdots \ge y_n & \text{when } b < 0 \end{array}\right\} \text{—— (ii)}$$

In either case, the middlemost value in (ii) corresponds to the middlemost value in (i). Hence

$$Me(y) = a + b\,Me(x).$$

The most generalised result in the connection states that, if $y = h(x)$ be a monotonic function of x, then

$$Me(y) = h\{Me(x)\}.$$

C.U

Quantiles (fractiles) :→ The p^{th} quantile of a set of values of a variable is that value of the variable below which we have 'p' or more than 'p' proportion of values and above which we have $(1-p)$ or more than '$1-p$' proportion of values.

Hence, for n values of a variable, the p^{th} quantile (ξ_p) is that value of the variable for which

$\{$No of values which are $\le \xi_p\} \ge np$

and $\{$No. of values which are $\ge \xi_p\} \ge n(1-p)$

Quartiles : If $p = \frac{1}{4}$, then $\xi_{1/4}$ is called the 1^{st} quartile. Similarly $\xi_{1/2}$, $\xi_{3/4}$ are called the 2^{nd} quartile (median), 3rd quartile.

Deciles : $\xi_{1/10}, \xi_{2/10}, \ldots, \xi_{9/10}$ are known as the deciles which divide the whole frequency distribution into ten equal parts.

C.U

Percentiles: $\xi_{1/100}, \xi_{2/100}, \ldots\ldots, \xi_{99/100}$ are known as the percentiles

Example: In a set of 11 or 12 values, find the 1st quartiles.

C.U **Solution:** Let the set contains 11 values $x_1, x_2, \ldots\ldots, x_{11}$ of a variable x.

Let us consider the ordered values

$$x_{(1)} \le x_{(2)} \le x_{(3)} \le \ldots\ldots \le x_{(11)}$$

By Def^n.,

$$\{\text{No. of values } x_i \le \xi_{1/4}\} \geqslant 11 \times \frac{1}{4}$$

and $\{\text{No. of values } x_i \geqslant \xi_{1/4}\} \geqslant 11\left(1 - \frac{1}{4}\right)$

Note that

$$\{\text{No. of values } x_i \le x_{(3)}\} = 3 = \frac{12}{4} > \frac{11}{4}$$

and $\{\text{No. of values } x_i \geqslant x_{(3)}\} = 9 = 12 \times \frac{3}{4} > 11 \times \frac{3}{4}$

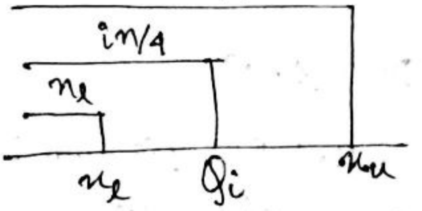
By def^n. $\xi_{1/4} = x_{(3)}$.

Continuous Variable ⇒ The frequency dist^n of a continuous variable needs special attention and the quartile Q_i

C.U

may be supposed to be the value for which the cumulative frequency is $i.n/4$, $i = 1, 2, 3$. By going through the $(\le)$-type cumulative frequency table, we can determine in which class interval the quartiles lie. Let us denote the lower and upper class boundaries of the class containing the quartile Q_i by x_l and x_u and the corresponding cumulative frequencies by n_l and n_u. The cumulative frequency of the ith quartile Q_i is $i.n/4$ and therefore, $(i.n/4 - n_l)$ is the frequency between the lower boundary (x_l) of the quartile class and Q_i. Assuming the frequencies are uniformly distributed over the quartile class, we have

$$\frac{n_u - n_l}{x_u - x_l} = \frac{i.\frac{n}{4} - n_l}{Q_i - x_l} = \text{the frequency per unit length}$$



Hence, $\dfrac{Q_i - x_l}{x_u - x_l} = \dfrac{i\frac{n}{4} - n_l}{n_u - n_l}$

$\Rightarrow \dfrac{Q_i - x_l}{c} = \dfrac{in/4 - n_l}{f}$, where c and f are the width and frequency of the quartile class.

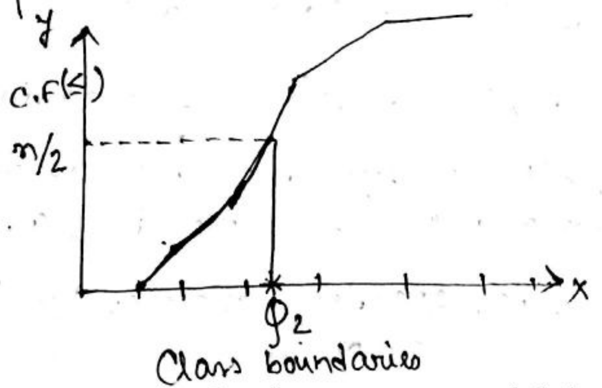
$$\Rightarrow Q_i = x_l + \frac{\left(i\frac{n}{4} - n_l\right)}{f} \times c, \; i = 1, 2, 3$$

In particular,

$$Q_2 \text{ (the median)} = x_l + \frac{\frac{n}{2} - n_l}{f} \times c$$

Graphical method of finding Quartiles:

Define $g(x) = \{\text{No. of values } x_i \le x\}$.
Then the graph $y=g(x)$ is called the ogive of less than type.
In grouped data,



In grouped data, c.f of $\le$ type are plotted for different class boundaries and the plotted points are then joined by straight line. (which is equivalent to the assumption that frequencies are uniformly distributed over the classes).

By defn. $g(Q_i) = i.\frac{n}{4}$, $i=1,2,3$.

To determine the value of the quartile Q_i, we mark a point along the y-axis corresponding to $i.\frac{n}{4}$, $i=1,2,3$ and from this pt. draw a line parallel to the x-axis and find the pt. where this line cuts the curve $y=g(x)$, draw a line perpendicular to x-axis from this pt. The distribution between the origin and the foot of the perpendicular is the quartile Q_i.

Remark:→ The p^{th} quantile of a grouped freq.distn. is given by

$$\xi_p = x_l + \frac{np - n_l}{f} \times c.$$

$\boxed{\text{C.U}}$

Example:→ Find the 1st and 3rd quartiles in a set of 11 values.

Soln.:→ The pth order quantile of a data is the value (ξ_p) such that

$$\left\{\frac{\text{No. of values which are} \le \xi_p}{n}\right\} \ge p$$

and

$$\left\{\frac{\text{No. of values which are} \ge \xi_p}{n}\right\} \ge 1-p$$

$\xi_{1/4}$ → 1st quartile, $\xi_{1/2}$ → 2nd quartile, $\xi_{3/4}$ → 3rd quartile,
$\xi_{1/10}$ → 1st decile, $\xi_{1/100}$ → 1st percentile.

Let $x_1 < x_2 < x_3 < \cdots < x_n$ be 11 values of a variable.

Note that,

$$\frac{\{\text{No. of values which are} \le x_i\}}{11} = \frac{3}{11} > \frac{3}{12} = \frac{1}{4}$$

1) How, in your opinion, should an average change when all values of the variable are increased or decreased - i) by the same amount? ii) in the same proportion? 4.3.

Soln :→ Let $x_1, x_2, \ldots, x_n$ be n values of a variable x.

i) Let $y_i = a + x_i$, i.e. all values are increased or decreased by the same amount, where $i = 1(1)n$.

A.M. → $\bar{y} = a + \bar{x}$, i.e. AM also increases by the same amount.

Median :→ Median$(y) = a +$ median(x), i.e. Median also changes by the same amount.

Mode :→ Mode$(y) = a +$ mode(x), i.e. Mode also changes by the same amount.

H.M. →
$$\text{H.M.}(y) = \frac{n}{\sum_i \frac{1}{y_i}}$$
$$= \frac{n}{\sum_i \frac{1}{a+x_i}} \neq a + \text{H.M}(x)$$

G.M. →
$$\text{G.M}(y) = \left(\prod_{i=1}^{n} y_i\right)^{1/n}$$
$$= \left\{\prod_{i=1}^{n} (a+x_i)\right\}^{1/n}$$
$$\neq a + \text{G.M}(x)$$

i.e. G.M. and H.M do not change by the same amount.

ii) Let, $y_i = bx_i \;\forall\; i = 1(1)n$, i.e. all values are increased or decreased in the same proportion.

Then, for, AM → $\bar{y} = b\bar{x}$; H.M. → HM$(y) = b$ H.M(x); GM$(y) = b$ GM(x); Median$(y) = b$ median(x); Mode$(y) = b$ mode(x).

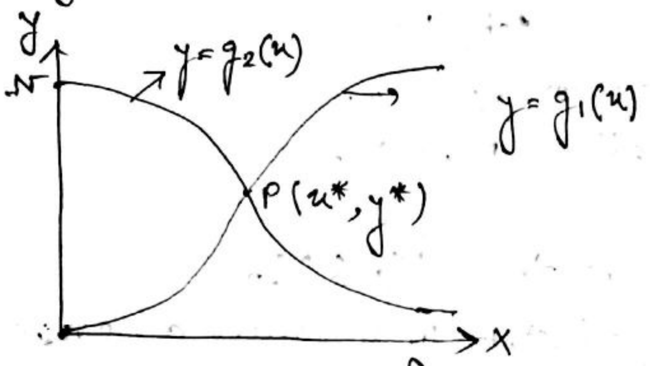
So, their averages also increased or decreased in the same proportion.

Scanned by CamScanner

(C.U)

2) Show that the median of a variable is the abscissa of the point of intersection of its two ogives (of the 'less than' & 'greater-than' type). 4.6.

Soln :->



Define, $g_1(x) = \{$ the number of values which are $\leq x\}$

$= \{$ cumulative frequencies of $\leq$ type of the value $x\}$

$g_2(x) = \{$ the number of values which are $\geq x\}$

$= \{$ cumulative frequencies of $\geq$ type of the value $x\}$

The graphs of $y = g_1(x)$ and $y = g_2(x)$ are the ogives of 'less than type' and 'greater than type' respectively.

Let $P(x^*, y^*)$ be the point of intersection of $y = g_1(x)$ and $y = g_2(x)$ then we have

$$g_1(x^*) = y^* = g_2(x^*)$$

$$\Rightarrow g_1(x^*) = g_2(x^*)$$

$\Rightarrow \{$ the number of values which are $\leq x^*\} = \{$ the number of values which are $\geq x^*\}$

Hence, x^*, the abscissa of the point of intersection of the two ogives, is the median.

Cor. The median is the x-value of the point of intersection of two ogives.

Soln :-> The two ogives ('less than' and 'greater than' types) intersect at a point which has y-value $\frac{N}{2}$. If not, there will be different proportion of observation below and above the point in the y-axis, which contradicts the point as the point of intersection.

The x-value corresponding to y-value $\frac{N}{2}$ is the median (by defn).

3) (a) There are two set of values of x. The first set with n_1 values has median M_1 and the second with n_2 values has median M_2. Show that the median of all n_1+n_2 values taken together must lie between M_1 and M_2. 4.7.(a)

C.U

Soln. :→

	1st set	2nd set
No. of values	n_1	n_2
Median	M_1	M_2

Let us assume that $M_1 \leq M_2$

From the 1st set, the no. of values which are less than equal to M_1 is $\frac{n_1}{2}$.

From the combined data, the no. of values which are less than equal to M_1 is $\frac{n_1}{2}+x$, clearly $0 \leq x \leq \frac{n_2}{2}$

Thus no. of values which are less than equal to M_1 is $\leq \frac{n_1+n_2}{2}$.

From the 2nd set, the no. of values which are greater than equal to M_2 is $\frac{n_2}{2}$.

From the combined data, the number of values which are greater than equal to M_2 is $\frac{n_2}{2}+y$, clearly $0 \leq y \leq \frac{n_1}{2}$.

Thus, the number of values which are greater than equal to M_2 is $\leq \frac{n_1+n_2}{2}$.

Hence, from the combined data, the number of values which are $\leq M_2$ is $\geq \frac{n_1+n_2}{2}$. Therefore, in the combined data, the no. of values which are $\leq M_1$ is $\leq \frac{n_1+n_2}{2}$:

Hence, the median (M) of the combined data lies between M_1 and M_2.

If $M_1 \geq M_2$, then we have

$$M_1 \geq M \geq M_2 .$$

Alternative method:→

Soln. :→ Set 1: $x_{11}, x_{12}, \ldots, x_{1n_1}$

Set 2: $x_{21}, x_{22}, \ldots, x_{2n_2}$

let, For set1, the median is M_1, set 2, the median is M_2.

Case-I:→ $x_{2i} \geq M_1$ for every $i=1(1)n_2$; $M_1 \leq M_2$.

So, below M_1 there are $\frac{n_1}{2}$ observations and at least $\frac{n_2}{2}$ obsn. above M_1 and $\frac{n_1+n_2}{2}$ obsn. below M_2. Hence combined median (M) will be above M_1 and below M_2. i.e. $M_1 \leq M \leq M_2$.

Case-II:→ $x_{2i} \geq x_{n_1}$ ∀ $i=1(1)n_2$; By the above argument, $M_1 \leq M \leq M_2$.

Hence, M lies between M_1 and M_2.

(b) Show that if $\bar{x}_1$ and $\bar{x}_2$ are the means of the two sets, then the mean $\bar{x}$ of the combined set also lie between $\bar{x}_1$ and $\bar{x}_2$.

4.7.(b)

Soln. :→ $\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1+n_2}$

Let, $\bar{x}_1 < \bar{x}_2$

Now $\bar{x}_1 - \bar{x} = \bar{x}_1 - \left(\frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1+n_2}\right)$

$= \frac{n_2(\bar{x}_1 - \bar{x}_2)}{n_1+n_2} < 0$

i.e. $\bar{x}_1 < \bar{x}$.

and $\bar{x}_2 - \bar{x} = \bar{x}_2 - \left(\frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1+n_2}\right)$

$= \frac{n_1(\bar{x}_2 - \bar{x}_1)}{n_1+n_2} > 0$.

i.e. $\bar{x} < \bar{x}_2$

when $\bar{x}_1 < \bar{x}_2$, $\bar{x}_1 < \bar{x} < \bar{x}_2$

similarly when, $\bar{x}_1 > \bar{x}_2$, then $\bar{x}_1 > \bar{x} > x_2$

Hence, $\bar{x}$ lies between $\bar{x}_1$ and $\bar{x}_2$.

Alt. method:→

$\bar{x}_1 > \bar{x}_2$

$\therefore n_1\bar{x}_1 > n_1\bar{x}_2$

$\therefore n_1\bar{x}_1 + n_2\bar{x}_2 > n_1\bar{x}_2 + n_2\bar{x}_2$

$\Rightarrow \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1+n_2} > \bar{x}_2$.

$\Rightarrow \bar{x} > \bar{x}_2$. —①

$\bar{x}_1 > \bar{x}_2$

$\therefore n_2\bar{x}_1 > n_2\bar{x}_2$

$\therefore n_1\bar{x}_1 + n_2\bar{x}_1 > n_1\bar{x}_1 + n_2\bar{x}_2$

$\therefore \bar{x}_1 > \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1+n_2}$

$\Rightarrow \bar{x}_1 > \bar{x}$. —②

combining ① and ② we get, $\bar{x}_1 > \bar{x} > \bar{x}_2$ when $\bar{x}_1 > x_2$.

Hence, $\bar{x}$ lies between $\bar{x}_1$ and $\bar{x}_2$.

4) ** Let x be a variable assuming the values $1, 2, \ldots, k$ and let $F_1' = n, F_2', \ldots, F_k'$ be the corresponding cumulative frequencies of the 'greater-than' type, show that 4.8.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} F_i'$$

Soln :->

x	f	Cumulative frequency ($\geq$)
1	f_1	$n = F_1' = f_1 + \cdots + f_k$
2	f_2	$F_2' = f_2 + \cdots + f_k$
3	f_3	$F_3' = f_3 + \cdots + f_k$
$\vdots$	$\vdots$	$\vdots$
$k-1$	f_{k-1}	$F_{k-1}' = f_k + f_{k-1}$
k	f_k	$F_k' = f_k$
TOTAL =	n	

$$\begin{aligned}\sum_{i=1}^{k} F_i' &= F_1' + F_2' + \cdots + F_{k-1}' + F_k' \\ &= f_1 + f_2 + \cdots + f_{k-1} + f_k \\ &\quad + f_2 + \cdots + f_{k-1} + f_k \\ &\quad + f_3 + \cdots + f_{k-1} + f_k \\ &\quad \vdots \\ &\quad \cdots + f_{k-1} + f_k \\ &\quad + f_k \\ &= 1 \cdot f_1 + 2 \cdot f_2 + \cdots + (k-1) f_{k-1} + k \cdot f_k\end{aligned}$$

~~[illegible]~~ $= \sum_{i=1}^{k} x_i f_i$

$= n\bar{x}$

So, $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} F_i'$

Scanned by CamScanner

5)(a) Suppose x is a variable (discrete or continineuous) with median M_i. If $y = g(x)$ be a monotonically increasing or decreasing function of x , show that the median of y is $g(M_i)$? 4.9.(a)

Sol^n :→ Let $x_1, \ldots\ldots, x_m$ be 'n' values of a variable x.

i) Let $y = g(x)$ be a monotonically increasing function.

Then let $x_1 \le x_2 \le \ldots\ldots \le x_m \le \ldots\ldots \le x_n$ be the ordered values of x and x_m be the median.

Therefore, we have -

$$g(x_1) \le g(x_2) \le \ldots\ldots \le g(x_m) \le \ldots\ldots \le g(x_n)$$

and as x_m is the middle most value in ordered arrangement of values of x, then $g(x_m)$ is also the middle most values in the ordered arrangements of the values of $g(x)$.

Hence, $g(x_m)$ is the median of the values of $y = g(x)$, $\{m = [\frac{n}{2}] + 1\}$

$\therefore\ Me(y) = g(Me(x))$.

ii) Let $y = g(x)$ be decreasing.

then $g(x_1) \geq g(x_2) \geq \ldots\ldots \geq g(x_m) \geq \ldots\ldots \geq g(x_n)$

Below $x_{(m)}$ among values of x, we have half of the values.

Since, $g(x)$ is decreasing, above $g(x_{(m)})$ among the values of $g(x)$ we have half of the values.

Hence, $g(x_{(m)})$ is the median of the values of $g(x)$.

(b) Can a similar statement be made with regard to the mean? 4.9.(b)

Sol^n :→

Property of Mode :→

6) If $y=a+bx$ and M_0 is the mode of x, then show that the mode of y must be $a+bM_0$.

4.10

Soln. :→

① Values of x	② Frequency	③ Values of $y=a+bx$
x_1	f_1	$y_1=a+bx_1$
x_2	f_2	$y_2=a+bx_2$
⋮	⋮	⋮
M_0	f_m	$y_m=a+bM_0$
⋮	⋮	⋮
x_k	f_k	$y_k=a+bx_k$

The column ② and ③ constitute the frequency distribution of y. Here, the frequency f_m of the value $y_m=a+bM_0$ is the maximum frequency. Hence, the mode of y is

$$y_m = a+bM_0 .$$

$$\therefore \text{Mode}(y) = a+b.\text{Mode}(x)$$

Alternative Method :→ Let $x_1, x_2, \ldots, x_n$ be the n obsn. of x. Now, if we modify the obsn into its ascending order of magnitude then without loss of generality, we can write $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$.

Let $x_{(m)}$ be the M_0 of x. Then naturally $a+bx_{(1)}$ be $y_{(1)}$ as a and b are constant.

then $a+bx_{(n)}$ be $y_{(n)}$.

As $x_{(m)}$ be the mode of x then without loss of generality $y_{(m)}$ i.e. $a+b.x_{(m)}$ be the mode of y.

Scanned by CamScanner

7) Let x be a variable assuming positive values only. Show that-
(a) the A.M of the reciprocal of x can't be smaller than the reciprocal of its AM.

(b) the AM of the square root of x can't be greater than the square root of its AM.

4.12

Sol^n:→ Let $x_1, x_2, \ldots, x_n$ be n values of a positive variable x.

Cauchy Schwarz inequality -

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \le \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right)$$

(a) Take $a_i = \frac{1}{\sqrt{x_i}}$, $b_i = \sqrt{x_i}$, $i = 1(1)n$

Then from C-S inequality -

$$\left(\sum_{i=1}^{n} \frac{1}{\sqrt{x_i}} \cdot \sqrt{x_i}\right)^2 \le \left(\sum_{i=1}^{n} \frac{1}{x_i}\right)\left(\sum_{i=1}^{n} x_i\right)$$

$$\Rightarrow n^2 \le \left(\sum_{i=1}^{n} \frac{1}{x_i}\right)\left(\sum_{i=1}^{n} x_i\right)$$

$$\Rightarrow \frac{n}{\sum_{i=1}^{n} x_i} \le \frac{\sum_{i=1}^{n} \frac{1}{x_i}}{n}$$

$$\boxed{\cancel{\Rightarrow \frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i} \geq \frac{1}{\frac{\sum x_i}{n}}}} \quad \Rightarrow \frac{1}{\bar{x}} \le \overline{\left(\frac{1}{x}\right)}$$

$$\Rightarrow \overline{\left(\frac{1}{x}\right)} \nless \frac{1}{\bar{x}}.$$

(b) Take $a_i = \sqrt{x_i}$, $b_i = 1$, $\forall\ i = 1(1)n$.

Then, from C-S inequality,

$$\left(\sum_{i=1}^{n} \sqrt{x_i} \cdot 1\right)^2 \le \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} 1\right)$$

$$\Rightarrow \left(\sum_{i=1}^{n} \sqrt{x_i}\right)^2 \le \left(\sum_{i=1}^{n} x_i\right).n$$

$$\Rightarrow \left(\frac{\sum_{i=1}^{n} \sqrt{x_i}}{n}\right)^2 \le \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\Rightarrow \left(\overline{\sqrt{x}}\right)^2 \le \bar{x}$$

$$\Rightarrow \left(\overline{\sqrt{x}}\right) \ngtr \sqrt{\bar{x}}$$

8) For a frequency distribution the upper class boundary bears a constant ratio r to the lower class boundary. If x_i and f_i be respectively the classmark and the frequency of the ith class and G be the geometric mean of the distribution, show that

$$\log G = \log x_1 + \frac{\log r}{N}\sum_{i=1}^{k}(i-1)f_i, \quad \text{where } N=\sum f_i.$$

$\boxed{\text{C.U}}$ 4.14.

$\boxed{\text{Or}}$

In a frequency distribution the upper class boundary has a constant ratio to the lower class boundary. Show that —

$$\log G = x_0 + \frac{C}{N}\sum_{i}(i-1)f_i, \text{ where —}$$

x_0 = log of the mid-value of 1st class interval,

C = log of the ratio between upper boundary and lower boundary's.

Soln :→

Class-boundary	Class mark	Frequency
$x_0'-x_1'$	$x_1 = \frac{x_0'(r+1)}{2}$	f_1
$x_1'-x_2'$	$x_2 = \frac{rx_0'(r+1)}{2} = rx_1$	f_2
$x_2'-x_3'$	$x_3 = \frac{r^2x_0'(r+1)}{2} = r^2x_1$	f_3
$\vdots$	$\vdots$	$\vdots$
$x_{i-1}'-x_i'$	$x_i = \frac{r^{i-1}x_0'(r+1)}{2} = r^{i-1}x_1$	f_i
$\vdots$	$\vdots$	$\vdots$
$x_{k-1}'-x_k'$	$x_k = \frac{r^{k-1}x_0'(r+1)}{2} = r^{k-1}x_1$	f_k

$$\frac{x_1'}{x_0'} = r, \quad \frac{x_2'}{x_1'} = r, \; \cdots\cdots, \; \frac{x_i'}{x_{i-1}'} = r, \; \cdots\cdots, \; \frac{x_k'}{x_{k-1}'} = r.$$

$$x_1 = \frac{x_1'+x_0'}{2} = \frac{rx_0'+x_0'}{2} = \frac{x_0'(r+1)}{2}$$

$$x_2 = \frac{x_2'+x_1'}{2} = \frac{r^2x_0'+rx_0'}{2} = \frac{rx_0'(r+1)}{2} = rx_1$$

$$\vdots$$

$$x_i = \frac{x_i'+x_{i-1}'}{2} = \frac{r^ix_0'+r^{i-1}x_0'}{2} = \frac{r^{i-1}x_0'(r+1)}{2} = r^{i-1}x_1$$

$$\vdots$$

$$x_k = \frac{r^{k-1}x_0'(r+1)}{2}$$

$$\log G = \frac{1}{N}\sum_{i=1}^{k} f_i \log x_i$$

$$= \frac{1}{N}\sum_{i=1}^{k} f_i \log\left\{\frac{r^{i-1}.x_0'(r+1)}{2}\right\}$$

$$= \frac{1}{N}\sum_{i=1}^{k} f_i \log\left\{x_1 . r^{i-1}\right\}$$

$$= \frac{1}{N}\sum_{i=1}^{k} f_i\left\{\log x_1 + (i-1)\log r\right\}$$

$$= \log x_1 \cdot \frac{1}{N}\left(\sum_{i=1}^{k} f_i\right) + \frac{\log r}{N}\sum_{i=1}^{k}(i-1)f_i$$

$[\because N = \sum f_i]$

$$= \log x_1 + \frac{\log r}{N}\sum_{i=1}^{k}(i-1)f_i$$

$$= x_0 + \frac{c}{N}\sum_{i=1}^{k}(i-1)f_i$$

$\left[\text{where, } \log x_1 = x_0, \; \log r = c\right]$

9) If $a \le x_i \le b$ for $i = 1(1)n$, then prove that —
i) $a \le G \le b$, ii) $a \le H \le b$.

Soln :→ i) $\prod_{i=1}^{n} a \le \prod_{i=1}^{n} x_i \le \prod_{i=1}^{n} b$

or, $a^n \le G^n \le b^n$

$\therefore$ $a \le G \le b$, where n is always positive, n = even

ii) $a \le x_i \le b$

or, $\frac{n}{b} \le \sum_{i=1}^{n} x_i \le \frac{n}{a}$

or, $\frac{1}{b} \le \frac{1}{n}\sum x_i \le \frac{1}{a}$

or, $b \geqslant \frac{n}{\sum x_i} \geqslant a$

or, $a \le H \le b$.

10) $\dfrac{\sum_{i=1}^{k} x_i w_i}{\sum_{i=1}^{k} w_i} \geq \dfrac{1}{k}\sum_{i=1}^{k} x_i$ provided $w_i \gtreqless w_j$ according as $x_i \gtreqless x_j$.

$\boxed{\text{C.U}}$

or,

if Show that weighted AM can't be less than simple AM greated weights are associated to the higher values.

Soln :→

$$\sum_{i=1}^{k}\sum_{j=1}^{k}(x_i - x_j)(w_i - w_j) \geq 0$$

$$\text{or, } 2k\sum_i x_i w_i - 2\sum_i x_i \sum_j w_j \geq 0$$

$$\text{or, } k\sum_i x_i w_i - \sum_i x_i \sum_j w_j \geq 0$$

$$\text{or, } k\frac{\sum x_i w_i}{\sum w_j} - \sum x_i \geq 0$$

$$\text{or, } \frac{\sum x_i w_i}{\sum w_j} - \frac{1}{k}\sum x_i \geq 0$$

$$\text{or, } \overline{x_w} - \overline{x} \geq 0$$

$$\text{or, } \overline{x_w} \geq \overline{x}.$$

11) Show that ~~observations~~ AM ≥ GM ≥ HM. when does equality hold?

$\boxed{\text{C.U}}$

Soln. Let $x_1, x_2, \ldots, x_n$ be n observations of the variable $x \ni x_i > 0 \; \forall i$

$AM = \frac{1}{n}\sum x_i = A$ (say)

$GM = (\prod_i x_i)^{1/n} = G$ (say)

$HM = \dfrac{n}{\sum\left(\frac{1}{x_i}\right)} = H$ (say)

Lemma:- If $u_1, u_2, \ldots, u_n$ be n observation of the variable $u \ni \prod_i u_i = 1$ then $\sum_i u_i \geq n$.

Proof:→ First we prove that $\log u \leq u - 1$

Case I:- let $0 < u < 1$

then $\log u = \log(u - 1 + 1)$

$= (u-1) - \frac{(u-1)^2}{2} + \frac{(u-1)^3}{3} - \ldots$

$\log u \leq (u-1)$ [Since $(u-1)$ is negative fractional quantity]

Case II :- let $u \geqslant 1$

Define, $y = \log u - u + 1$

so, $\frac{dy}{du} = \frac{1}{u} - 1$

$\frac{dy}{du} = 0 \Rightarrow u = 1$

$$\frac{d^2y}{du^2} = -\frac{1}{u^2}$$
$$= -1 < 0 \text{ at } u = 1$$

hence y is max when $u = 1$

i.e. $y_{max} = \log(1) - 1 + 1$
$= 0$

$\therefore y \leq 0$

or, $\log u \leq u - 1$

$\therefore$ we can say $\log u \leq u - 1 \;\forall u$

In general, $\log u_i \leq u_i - 1$

$$\sum_i \log u_i \leq \sum_i u_i - n$$

$$\log(u_1 \cdots u_n) \leq \sum_i u_i - n$$

$$\sum_i u_i \geqslant n$$

Now, $G = (x_1 x_2 \cdots x_n)^{1/n}$

$G^n = x_1 x_2 \cdots x_n$

$$\frac{x_1 \cdots x_n}{G^n} = 1.$$

By proposition, $\sum_i \frac{x_i}{G} \geqslant n$

$$\frac{1}{n}\sum_i x_i \geqslant G$$

i.e. $A \geqslant G$ —— ①

Again, $G^n = x_1 \cdots x_n$

$$\frac{G^n}{x_1 \cdots x_n} = 1$$

By proposition, $\sum_i \frac{G}{x_i} \geqslant n$

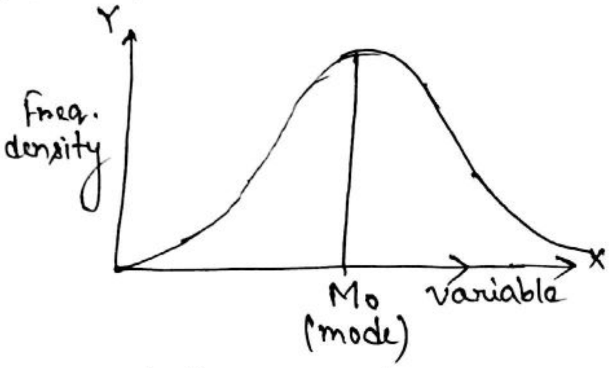
$$\frac{n}{\sum_i \frac{1}{x_i}} \leq G$$

i.e. $H \leq G$ —— ②

By ① & ②, we can say $A \geqslant G \geqslant H$

MODE :→ Mode is the value which occurs most frequently in a set of obs^{n}s. i.e. mode is the value of the variable which is pre-dominant in the series. Hence the mode of a variable is the value of the variable having the highest frequency. This def^n applies to a discrete variable only and for a discrete variable, the mo decan be immediately found by inspection.

For a continuous variable, the mode is the value of the variable with the highest frequency density corresponding to the ideal dist^n which would be obtained if the total freq. tends to infinity and if the class-width tends to zero i.e. it may be looked upon as the abscissa corresponds to the highest ordinate in the freq. curve. Mode is sometimes denoted by $\breve{x}$.



A-Frequency Curve

Discuss the method of obtaining mode for grouped data or Continuous variable when the class-widths are equal.

$\boxed{\text{C.U}}$

Ans:→ Let x_l and x_u be respectively the lower and upper boundaries of the modal class, and f_{m-1}, f_m and f_{m+1} the frequencies of the three classes. Since, in practice, we usually see that class frequencies, starting from a low value, gradually rise to a maximum and then, again, gradually come down, it can be expected that

$$M_0 - x_l \gtreqless x_u - M_0 \text{ according as } f_m - f_{m-1} \gtreqless f_m - f_{m+1},$$

M_0 being mode. Here it is assumed that

$$\frac{M_0 - x_l}{x_u - M_0} = \frac{f_m - f_{m-1}}{f_m - f_{m+1}}$$

From this we have,

$$\frac{M_0 - x_l}{(M_0 - x_l) + (x_u - M_0)} = \frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})}$$

or,

$$\frac{M_0 - x_l}{x_u - x_l} = \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}}$$

or, $M_0 = x_l + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \times c$

c being the common width, $c = x_u - x_l$.

Alternative Method :→

[Let us consider the grouped data :

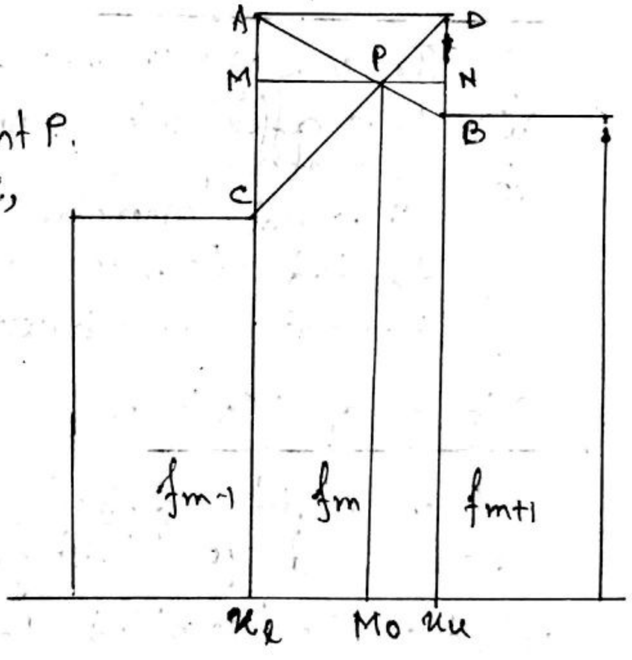
Class: $x_1 - x_2$ $x_2 - x_3$ .... $x_k - x_{k+1}$ ..... $x_{n-1} - x_n$

Freq: f_1 f_2 .... f_k .... f_{n-1}

where all classes have equal width 'c'.

If f_k is the maximum frequency, then the modal class is $x_k - x_{k+1}$]

Method - II.



The mode is the abscissa of the point P. where AB and CD intersect. PM ⊥ AC, PN ⊥ BD. From the geometry of similar triangles PAC and PBD, we have,

$$\frac{PM}{PN} = \frac{AC}{BD}$$

or, $$\frac{M_0 - x_l}{x_u - M_0} = \frac{f_m - f_{m-1}}{f_m - f_{m+1}}$$

This leads to,
$$M_0 = x_l + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \times c.$$

Method III.

Another method of determining mode is to use the empirical relation between mean, median and mode which is found to hold for unimodal distn. that do not deviate much from symmetry. The relation is

mean - mode = 3 (mean - median).

Symbolically, $M_0 = 3M_e - 2\bar{x}$.

⇨ Merits & Demerits of Median and Mode :→ C.U

⇨ Merits of Median :-

i) It is rigidly defined,

ii) It is based on all observations.

iii) It is readily comprehensible.

iv) It can be obtained without necessit y of measuring all the objects to be observed, in any case in which the objects can be arranged in order of magnitude.

v) It is least affected by abnormally large or small values of the variable.

⇨ Demerits of Median :-

i) 'In' case of even number of observations it is not rigidly defined.

ii) It is affected by sampling fluctuations except some specific situation.

iii) It is not amenable to algebric treatment./It can't be treated algebrically.

iv) It is not well-understood.

⇨ Merits of Mode :-

i) It is rigidly defined except when there are more than one value with the highest frequency or frequency density.

ii) It can be determined by inspection.

iii) It remains unchanged when some observations may be altered.

iv) It is easily comprehensible.

v) It can be determined if one or both of the terminal classes are open.

⇨ Demerits of Mode :-

i) Mode does not lend itself to algebric treatment.

ii) It is less reliable and less stable as regards sampling fluctuation.

iii) the determination of mode in the continuous case is impossible if only a few values of the variable are given.

Result :→ If $y = g(x)$ is a one-to-one (monotonic) function, then show that mode of $y = g(\text{mode of } x)$. 4.9.(b)

Proof :→ Consider a frequency distribution of the variate x:

Values	Frequency
x_1	f_1
x_2	f_2
$\vdots$	$\vdots$
$M(x)$	$f_m \to$ (maximum)
$\vdots$	$\vdots$
x_k	f_k
	n

If f_m is the max freq. then $M(x)$ is the mode of x.

The freq distn of y is:

Value of x	value of y	freq.
x_1	$y_1 = g_1(x)$	f_1
x_2	$y_2 = g_2(x)$	f_2
$\vdots$	$\vdots$	$\vdots$
$M(x)$	$M(y) = g(M(x))$	$f_m \to$ max
$\vdots$	$\vdots$	$\vdots$
x_k	$y_k = g(x_k)$	f_k

Note that, freq. of $y_i = g(x_i)$ = frequency of x_i, since $y = g(x)$ is one-to-one i.e. for every y_i, we have an one or unique x_i.

Note that, the freq. of $M(y)$ is f_m, which is the max-freq.

Hence, mode of $y = M(y) = g(M(x)) = g(\text{Mode of } x)$.

⇨ What are the desiderata of a good measure of central tendency? Compare the mean, the median and the mode in the light of these desiderata. 4.2.

Ans:→

i) Requirements of an ideal average :—

Yule suggested the following chief characteristics of an ideal measure of central tendency :

(a) It should be rigidly defined,
(b) It should be easily understandable and easy to calculate.
(c) It should be amenable to algebraic treatment.
(d) It should be least affected by fluctuations of sampling.
(e) It should be based on all the observations.
(f) It should not be affected much by extreme values.

ii) Comparison of mean, median and mode:—

The mean is rigidly defined. The median is also rigidly defined, except when there is an even number of obsns. The mode is too rigidly defined when there are more than one value with the highest frequency or frequency-density.

In finding each of these measures, all the observations are taken into consideration. However, only the mean is directly based on all the values and its value changes even if a single observation is altered. On the other hand, median and mode may remain unchanged even after the alternation of several observations.

The significance of all the measures is quite easily comprehensible.

In general, the labour involved in the computation of all the three measures is almost the same. But in practice, the exact determination of mode of a continuous variable is impossible, because practically we never get an ideal distribution (or the frequency curve).

The mean has several properties and by virtue of them it can be readily manipulated in varied situations. But the median and mode don't possess such desirable properties.

Among the three measures, the mean is generally found to be least affected by sampling fluctuations. However, in this respect, the median or mode may be better than mean, in some specific situations.

Thus it is evident that, in general, mean is the best measure of central tendency. But there are situations where mean can't be used or should not be used. In the case of a grouped frequency distribution, if one or both of the terminal classes are open, mean is indeterminate. Again, if there be a few extreme values markedly different from majority of the values, mean should not be used as an average. In such situations, median or mode would be appropriate measure.

➩ Give an example of a situation where median would be the appropriate average than the mean.

Am. → Let the weights of 8 iron balls be 138, 143, 141, 139, 152, 148, 160 and 267 kg. Here the mean is 161 kg, but this cannot be said to be a representative value, because the seven out of eight given value are smaller than 161, the mean value.

In case of this sort, Where the data contain a few extreme values widely different from the majority of the values, the mean should not be used. In the present example, the median would be appropriate average.

We are arranging this data in order of magnitude.

138, 139, 141, 143, 148, 152, 160, 267.

Here median = the AM of (4th value + 5th value)

= (143 + 148) / 2

= 145.

➩ Define a particular situation when mean is superior than median or mode.

Ans:→ In case of algebrical treatment.

⇨ The AM, GM, HM of 3 observations are 4, 3.63, 2.67 respectively. Find out the observations.

Ans:—

$$x+y+z=12 \quad —— \textcircled{1}$$

$$xyz = 48 \quad —— \textcircled{2}$$

$$\frac{3}{\frac{1}{x}+\frac{1}{y}+\frac{1}{z}} = 2.67$$

$$\Rightarrow xy+yz+zx = 44 \quad —— \textcircled{3}$$

$$\textcircled{1} \Rightarrow x = 12-(y+z) \Rightarrow (y+z) = 12-x$$

$$\textcircled{2} \Rightarrow x = \frac{48}{yz} \qquad \Rightarrow yz = \frac{48}{x}$$

$$\therefore \textcircled{3} \Rightarrow x(y+z)+yz = 44$$

$$\Rightarrow x(12-x)+\frac{48}{x} = 44$$

$$\Rightarrow 12x^2 - x^3 + 48 = 44x$$

$$\Rightarrow x^3 + 144x - 12x^2 + 48 = 0$$

$$\Rightarrow x = 2.$$

$$\therefore y = 4,\ z = 6 \ \left[\text{from } \textcircled{2} \text{ and } \textcircled{3}\right]$$

QUANTILES :— Quantities such as quartiles, deciles, etc, which divide the total frequency into a number of parts, are called quantiles or grades, and when we speak of the grade of an individual we mean thereby the proportion of the total frequency which lies below it. Conventionally, half the individual is regarded as lying above, and half below, the point determined by the variate value which it bears. By quartiles the total frequency is divided into 4 parts., we may divide it into 100 parts by what are called percentiles, we may also divide into 10 parts by deciles.

⇨ Other measures of Central Tendency :—

[c.u.]

i) **Geometric Mean :→** The geometric mean of a set of n values of a variable is the n th root of their product. If a variable x assumes n values $x_1, x_2, \ldots, x_n$, then its geometric mean, denoted by x_g, is

$$x_g = (x_1 x_2 \cdots\cdots x_n)^{1/n} = \left(\prod_{i=1}^{n} x_i\right)^{1/n}$$

For a frequency distribution,

$$x_g = \left(\prod_{i=1}^{n} x_i^{f_i}\right)^{1/n}, \quad n = \sum_{i=1}^{n} f_i .$$

⇨ Properties of G.M. →

(a) If the given values of a variable are all equal, then the GM will be equal to their common value.

Suppose each of the n values of a variable x is equal to c.

Then $\prod_{i=1}^{n} x_i = \prod_{i=1}^{n} c = c^n$.

Hence $x_g = \left(\prod_{i=1}^{n} x_i\right)^{1/n} = (c^n)^{1/n} = c$.

(b) The logarithm of the GM of a set of values of a variable is the AM of their logarithms.

Suppose we have n values $x_1, x_2, \ldots, x_n$ of variable x and x_g denotes their geometric mean. then

$$x_g = \left(\prod_{i=1}^{n} x_i\right)^{1/n}$$

Taking logarithm of both sides, we have

$$\log x_g = \log\left(\prod_{i=1}^{n} x_i\right)^{1/n}$$

$$\text{or, } \log x_g = \frac{1}{n}\sum_{i=1}^{n} \log x_i$$

(c) If y is a function of a variable x in the form $y = ax$, then the GM of y is related to that of x in the similar form.

Let x_g and y_g be the GM of the variables x and y respectively and n be the number of given values.

Here $y_i = ax_i$, for each i,

so that, $\prod_{i=1}^{n} y_i = \prod_{i=1}^{n} (ax_i)$

$$= a^n \prod_{i=1}^{n} x_i \, .$$

$$\therefore \left(\prod_{i=1}^{n} y_i\right)^{1/n} = a\left(\prod_{i=1}^{n} x_i\right)^{1/n}$$

$$\therefore \; y_g = a x_g \, .$$

(d) The GM of the ratio of two variables is the ratio of their GMs.

Suppose x and y are two variables and their values are given for each of n individuals. Let $u = \frac{x}{y}$. If x_i and y_i denote the values of x and y for ith individual, $i = 1, 2, \ldots, n$ then

$$u_i = \frac{x_i}{y_i} \quad \text{for each } i,$$

so that $\prod_{i=1}^{n} u_i = \prod_{i=1}^{n} (x_i/y_i) = \dfrac{\prod_{i=1}^{n} x_i}{\prod_{i=1}^{n} y_i}$

or, $\left(\prod_{i=1}^{n} u_i\right)^{1/n} = \left(\prod_{i=1}^{n} x_i \Big/ \prod_{i=1}^{n} y_i\right)^{1/n}$

$$= \left(\prod_{i=1}^{n} x_i\right)^{1/n} \Big/ \left(\prod_{i=1}^{n} y_i\right)^{1/n}$$

or, $u_g = x_g / y_g$.

Use of G.M. :→ In determining the average annual percentage rate of nit profit to sales of a company over a ten year period. – Here G.M. is appropriate

As GM of the ratio of two variables is the ratio of their GMs, it is sometimes preferred for averaging ratios of two variables, rates of population growth, rates of interest, rates of depreciation, etc.

GM is also used for finding the value of a variable at the mid point of a time period when the variable is an exponential function of time.

(e) If a variable (x) changes over time (t) exponentially, then the value of the variable at the mid-point of an interval (t_1,t_2) i.e. at $\frac{t_1+t_2}{2}$ is the GM of its values at t_1 and t_2.

Suppose $x_t = ab^t$.

Then $x_{t_1} = ab^{t_1}$ and $x_{t_2} = ab^{t_2}$.

Also, the value of x at $\frac{t_1+t_2}{2}$ is

$$ab^{\frac{t_1+t_2}{2}} = \{a^2 b^{(t_1+t_2)}\}^{1/2} = \{(ab^{t_1})(ab^{t_2})\}^{1/2} = (x_{t_1} \cdot x_{t_2})^{1/2}.$$

Advantages :->

i) The GM is rigidly defined.

ii) It is directly based on all obsns.

iii) It possesses some properties which enable the measure to be readily applicable in different situations.

iv) Generally, the presence of a few extremely small or large values has no considerable effect on GM.

Disadvantages :->

i) The GM is abstract in nature.

ii) It is difficult to compute.

iii) If a single value of a variable is zero, then the GM becomes zero, irrespective of the magnitudes of the other values. Also, it may be imaginary if some values are negative. Generally use of GM is restricted to positive values.

C.U

[redacted] iii) Harmonic Mean :-> The harmonic mean of a set of non-zero values of a variable is the reciprocal of the AM of the reciprocals of the values.

Thus, H.M. of n non-zero values $x_1, x_2, \dots, x_n$ of a variable x is

$$x_h = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}.$$

For a frequency distribution,

$$x_h = \frac{n}{\sum_{i=1}^{n} \frac{f_i}{x_i}}, \text{ where } n = \sum_{i=1}^{n} f_i$$

Use of H.M. :->

The HM is not commonly used, but it is the appropriate average when the variable is of the form "x per unit y", and equal amounts of x are considered. If, however, equal amounts of y are considered, AM is the appropriate average. — Two situations where AM & HM are appropriate.

⇨ Give example of the situations where AM and HM are appropriate.

Or,

C.U Suppose a train moves n equal distances each of s kms., say, with speeds $v_1, v_2, \ldots, v_n$ kms. per hour. Also suppose that the another train moves for n equal time intervals, each of length t hours, say, with the above speeds. Obtain the average speed in each case.

Ans:→ Case-I.

$$\text{Average speed} = \frac{\text{total distance}}{\text{total time}} = \frac{ns}{\frac{s}{v_1}+\frac{s}{v_2}+\cdots+\frac{s}{v_n}}$$

$$= \frac{n}{\frac{1}{v_1}+\frac{1}{v_2}+\cdots+\frac{1}{v_n}} \text{ km/hr.}$$

which is the HM of the given speeds.

Case-II.

$$\text{Average speed} = \frac{\text{total distance}}{\text{total time}} = \frac{v_1t+v_2t+\cdots+v_nt}{nt}$$

$$= \frac{v_1+v_2+\cdots+v_n}{n} \text{ km/hr.}$$

which is the AM of the given speeds.

⇨ Properties of H.M :→

(a) If the given values of a variable are all equal (≠0), then the HM will be equal to their common value.

Let n values of a variable x be given, each of which is equal to c (≠0).

Hence the HM of the values $= \frac{n}{\sum_{i=1}^{n}\frac{1}{c}} = \frac{n}{\frac{n}{c}} = c$.

(b) If a variable y is related to another variable x in the form y=ax, then the HM of y is related to that of x in the similar form.

Suppose n denotes the number of given values. If x_h and y_h be the HMs of the variables x and y respectively, then

$$y_h = \frac{n}{\sum_{i=1}^{n}\frac{1}{y_i}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{ax_i}} = \frac{n}{\frac{1}{a}\sum_{i=1}^{n}\frac{1}{x_i}} = a.x_h$$

(c) Suppose there are two sets of values of a variable x consisting of n_1 and n_2 values and their respective H.M's are H_1 and H_2; then the HM (H) of the combined set is given by

$$H = \frac{n_1+n_2}{\frac{n_1}{H_1}+\frac{n_2}{H_2}}.$$

Let x_{1i}, $(i=1,2,\dots,n)$ be the values of x in the first set and x_{2i} $(i=1,2,\dots,n)$ be those in the second set.

Then $H_1 = \dfrac{n_1}{\sum_{i=1}^{n_1}\frac{1}{x_{1i}}}$ and $H_2 = \dfrac{n_2}{\sum_{i=1}^{n_2}\frac{1}{x_{2i}}}$

Hence $\sum_{i=1}^{n_1}\frac{1}{x_{1i}} + \sum_{i=1}^{n_2}\frac{1}{x_{2i}} = \frac{n_1}{H_1}+\frac{n_2}{H_2}$, which is the sum of reciprocals of values in the combined set.

Since the total number of values in the combined set is (n_1+n_2), the HM (H) of this set is given by

$$H = \frac{n_1+n_2}{\frac{n_1}{H_1}+\frac{n_2}{H_2}}.$$

⇨ Advantages of HM :→

i) It is rigidly defined,

ii) It is directly based on all the values.

⇨ Disadvantages :→

i) It is undefined even if a single value is zero.

ii) It is abstract in nature.

iii) It involves a lot of competational labour.

iv) It is not amenable to algebric treatment.

⇨ Give two examples where the GM or the HM would be the appropriate type of average. 4.5. C.U

Ans :→ The answer has been discussed before, but here we will give examples

Example of GM :→ The ratios of the prices in 1994 to those in 1982 for four commodities are 0.92, 1.25, 1.75 and 0.85. To get the the average price-ratio we use GM.

$$\log x_g = (\log 0.92 + \log 1.25 + \log 1.75 + \log 0.85)/4$$

$$= 0.05829 = \log 1.1436$$

$$\therefore x_g = 1.144.$$

Example of HM :→ Suppose milk is sold at the rate of 1.80, 2.00, 2.20 and 2.50 rupees per litre in four different months. Assuming that equal amounts of money are spent on milk by a family in four months, the average price in rupees per lit. will be the HM of the given figures, i.e.

$$x_h = \frac{4}{\frac{1}{1.80}+\frac{1}{2.00}+\frac{1}{2.20}+\frac{1}{2.50}} = \frac{4}{1.091} = 2.091.$$

Weighted Mean :-> Weighted means are used when a proper weightage is given to each value. thus a set of weights $w_1, w_2, \dots, w_n$ are taken along with the values $x_1, x_2, \dots, x_n$ of a varialele x, where each weight represents the relative importance of the corresponding value in the given content.

The weighted arithmatic mean ($\bar{x}_w$) is expressed as

$$\bar{x}_w = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}.$$

The weighted geometric mean (x_{gw}) is expressed as

$$x_{gw} = \left(\prod_{i=1}^{n} x_i^{w_i}\right)^{1/\sum_{i=1}^{n} w_i}$$

the weighted harmonic mean (x_{hw}) is expressed as

$$x_{hw} = \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{w_i}{x_i}}.$$

Distinguish between trimmed mean and winsorised mean.

Ans:-> In statistical data, we generally use simple means and weighted means as the measure of location. In recent times the use of a modified arithmatic mean to eliminate the effect of a few extreme observations has come into use. It has been suggested that all values lower than the first quartile and all values higher than the third quartile be ignored and AM be calculated on the basis of the remaining observations only. Alternatively, one may replace each value lower than the first quartile by the value of the first quartile and each value higher than the third quartile by the value of the third quartile and compute the AM based on the modified set of observations. the mean obtained in the first case is called the trimmed mean and the one obtained in the second case is reffered to as the the Winsorised mean.

➪ Describe the concept of Trimmed mean & Winsorized Mean?

ANS:-

Trimmed Sample Mean:-

(a) Remove all observations below the first quartile. Remove all observations above the third quartile.

(b) Calculate the mean of the remaining observations.

Winsorized Sample Mean:-

(a) Replace each observation below the first quartile with the value of the first quartile. Replace each observation above the third quartile with the value of the third quartile. All other observations remain unchanged.

(b) Calculate the the mean of all the observations thus modified.

Both the trimmed and Winsorized means are undisturbed by the presence of a small fraction of unusual or erroneous observations that are extremely small or large. The first measure "trims off" the observations that lie outside the center half, the second moves these observations over to the corresponding quartile before averaging.

- **Trimmed Mean:-** The AM is highly affected by extreme values in the sample. To eliminate the effect of few extreme observations, a modified AM can be used.

Definition:- $\boxed{\text{p-trimmed mean}}$:⇝ Let $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$ be the observed sample. the trimmed mean with proportion p trimmed off from each end is denoted by $T(p)$ and is defined as

$$T(p) = \frac{1}{n(1-2p)} \sum_{i=np+1}^{n-np} x_{(i)} ,$$ provided np is an integer.

A trimmed mean is identified by the proportion that is trimmed off from each end of the ordered sample. Thus 20%-trimmed mean of a sample of size 10 is the sample mean of the six observations remaining after trimming off the two largest and 2 smallest observations.

Example:- Consider the ordered data:

−44, −2, −1, 16, 20, 40, 42, 44, 50, 55

∴ The sample mean is $\overline{x} = T(0.00) = \frac{1}{10}\sum_{i=1}^{10} x_{(i)} =$ ____.

∴ 20% trimmed mean is $= T(0.20) = \frac{1}{6}\sum_{i=3}^{8} x_{(i)} =$ ____.

Trimmed means are very stable and useful.

CHAPTER - 3

Measure of Dispersion

⇨ Define Dispersion :-> C.U

5.1.

$\boxed{1}$ Def^{n} :-> The spreadness or scatteredness of a set of obs^{n}. (among or from the average) is called dispersion.

By the word 'dispersion' we mean the 'degree of scatteredness', so by 'dispersion' we measure the 'degree of heterogenity' among the obs^{n} i.e. how far the obs^{n} are dispersed with respect to their measure of central tendency. Thus by 'dispersion' we measure the 'degree of variability' among the obs^{n} about their average. This feature of frequency distribution which represents the variability of given values or reflects how scattered the values are, is called its dispersion. It may be said that while the central tendency of a variable is the tendency of its values to be similar, its dispersion represents the tendency of the values to be different.

$\boxed{2}$ Ex :-> Temperature in

City A	City B
10	0
12	4
8	8
9	8
6	14
4	12
8	11

Both the data have the same average 8.1 but in the case of B, the scatter around the avg. is much greater than that in the case of city A. The avg. does not enable us to draw proper idea about the overall nature of the two data sets. Hence, it is necessary, besides mentioning the avg. value, to state how scattered the given values are among themselves.

⇨ Measures of dispersion :-> There are some common measures of dispersion :

5.1.

i) Range (R).

ii) Mean deviation (MD).

iii) Standard deviation (SD).

iv) Quartile deviation (QD).

— These are the absolute measures of dispersion.

⇨ Characteriotics of a good measure of dispersion :→

i) rigidly defined,

ii) easily comprehensible and easy to calculate.

iii) based on all obs^n..

iv) Amenable to further mathematical treatment.

⇨ Range :- $\boxed{\text{C.U}}$

The range of a variable is the simplest measure of its dispersion and it is defined as the difference between the greatest and the least of its given set of values.

It should be noted that if the data are given in a grouped frequency distribution, the range can be considered as the difference between the largest upper boundary and the smallest lower boundary.

Let $x_1, x_2, \ldots\ldots, x_n$ be n values of a variable x.

Define, range $= \max\limits_{1\le i\le n}\{x_i\} - \min\limits_{1\le i\le n}\{x_i\}$ as a measure of dispersion

□ Ques:→ Discuss the effect of a linear transformation $y = a+bx$ on Range / Discuss the effect of change of origin and scale on the range of a variable / If $y = a+bx$, then Range(y) = $|b|$Range(x). S.2.

Sol^n.→

Let us consider that x_{max} and x_{min} respectively denote the maximum and minimum of given values of x, while the corresponding values of y are y_{max} and y_{min}.

Hence, Case-I → for $b>0$, we have

$y_{max} = a + bx_{max}$,

and $y_{min} = a + b\ x_{min}$,

so that, $y_{max} - y_{min} = b(x_{max} - x_{min})$

i.e. Range(y) = bRange(x) ——— (1)

Case-II :→ for $b<0$, we get

$$y_{max} = a + b x_{min}$$

and $y_{min} = a + b x_{max}$

so that, $y_{max} - y_{min} = -b(x_{max} - x_{min})$

i.e., $\text{Range}(y) = -b\,\text{Range}(x)$. ——— (ii)

Combining (i) and (ii), we get

$$\boxed{\text{Range}(y) = |b|\,\text{Range}(x)}$$

⇨ Ques :→ What are the defects of the measure R? $\boxed{\text{C.U}}$

Ans:→ i) It is easy to calculate but it is not based on all the observations. ii) It is affected by the change of origin.

iii) It is affected by sampling flactuations.

iv) It is not amenable to algebric treatment.

— It does not posses the reauirities of a good measure of dispersion.

⇨ MEAN DEVIATION :→

Let A be the chosen average of the variable x. then $(x_i - A)$ is the deviation of x_i from the average A. Clearly, higher the deviations $(x_i - A)$, $i=1(1)n$ in magnitude, the higher is the dispersion. To get an idea of dispersion of the data we have to combine all the deviations. But $\frac{1}{n}\sum_{i=1}^{n}(x_i - A)$ can't serve the purpose, since it may be small even when individual deviations $x_i - A$ are large. To avoid this difficulty we should combine the deviation after ignoring their signs, which can be done in two ways:

(1) Here, we shall combine the absolute deviations $|x_i - A|$ and define,

Mean deviation about A = Mean of the absolute deviation about A $= MD_A = \frac{1}{n}\sum_{i=1}^{n}|x_i - A|$.

when $A = \bar{x}$, define,

Mean deviation about mean $= MD_{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$

It may be mentioned that mean deviation is generally calculated about the AM.

Again, if $x_1, x_2, \dots, x_n$ are the given values of a variable x and $f_1, f_2, \dots, f_n$ are the corresponding frequencies, then

$$MD_A = \frac{1}{N}\sum_{i=1}^{n} f_i|x_i - A| \quad , \text{where } N = \sum_{i=1}^{n} f_i .$$

(b) Here, to get a measure of dispersion, we shall combine the square deviations $(x_i - A)^2$, $i = 1(1)n$ instead of combining $|x_i - A|$.

Define, mean square deviation about A $= M.S.D_A = \frac{1}{n}\sum_{i=1}^{n}(x_i - A)^2$.

Define, root-mean-square deviation about A $= RMSD_A = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - A)^2}$,

The square root is taken in order to express the measure in the unit as those of x.

☐ Properties of MD :→

If two variables x and y are related as $y = a + bx$ —①

then $MD(y)_{A(y)} = |b| MD(x)_{A(x)}$, S.2

where $A(x)$ and $A(y)$ are corresponding value of x and y satisfying the given relation ①.

~~where A(x) and A(y) are corresponding value of x and y~~

[For e.g., when $A(x) = \bar{x}$, $A(y) = a + b\bar{x}$]

Here $y = a + bx$. Suppose y takes the value y_i when $x = x_i$, for $i = 1, 2, .., n$.

$\therefore \; y_i = a + bx_i$, for each i.

Again, $A(y) = a + bA(x)$.

Then $y_i - A(y) = b\{x_i - A(x)\}$

or, $|y_i - A(y)| = |b|\,|x_i - A(x)|$

Hence $\frac{1}{n}\sum_{i=1}^{n}|y_i - A(y)| = |b|\frac{1}{n}\sum_{i=1}^{n}|x_i - A(x)|$.

i.e. $MD(y)_{A(y)} = |b| MD(x)_{A(x)}$

Result :→ $MD_{\bar{x}} = \frac{1}{n}\sum |x_i - \bar{x}| = \frac{2}{n}\sum_{x_i > \bar{x}} (x_i - \bar{x}) = \frac{2}{n}\sum_{x_i < \bar{x}} (\bar{x} - x_i)$

Proof :→

$$\sum_{i=1}^{n} | x_i - \bar{x} |= 0$$

$$\Rightarrow \sum_{x_i < \bar{x}} (x_i - \bar{x}) + \sum_{x_i > \bar{x}} (x_i - \bar{x}) = 0$$

$$\Rightarrow \sum_{x_i > \bar{x}} (x_i - \bar{x}) = \sum_{x_i < \bar{x}} (\bar{x} - x_i)$$

$$M.D_{\bar{x}} = \frac{2}{n}\sum_{x_i < \bar{x}} (\bar{x} - x_i) = \frac{2}{n}\sum_{x_i > \bar{x}} (x_i - \bar{x})$$

therefore, it is enough to consider only one of the two sets of values of x in computing $MD_{\bar{x}}$.

STANDARD DEVIATION :

C.U. Ques. What is standard deviation? 5.4.

Ans:→ The standard deviation is the positive square-root of the arithmatic mean of the squares of all deviations, deviations being measured from the arithmatic-mean of the observation. We define the standard deviation $= s = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2}$. as a measure of dispersion.

If the given data are arranged in a frequency table, the S.D is given by

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{n} (x_i - \bar{x})^2 f_i} \text{ , where } N = \sum_{i=1}^{n} f_i .$$

Standard Deviation and its computations :→

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

For computational purpose this formula may be expressed in a simpler form. We have

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_i x_i^2 - 2\sum_i x_i \cdot \bar{x} + n\bar{x}^2$$
$$= \sum_i x_i^2 - n\bar{x}^2 , \left[\text{since } \sum_i x_i = n\bar{x}\right]$$

Hence, $s = \sqrt{\frac{1}{n}\sum_i x_i^2 - \bar{x}^2} = \sqrt{\frac{1}{n}\sum_i x_i^2 - \left(\frac{1}{n}\sum_i x_i\right)^2}$

For grouped data, the 's' is given by

$$s = \sqrt{\frac{1}{N}\sum_i x_i^2 f_i - \bar{x}^2} = \sqrt{\frac{1}{N}\sum_i x_i^2 f_i - \left(\frac{1}{N}\sum_i x_i f_i\right)^2}$$

It should be noted that s^2, the square of the standard deviation, is called the variance of the variable.

$$s^2(x) = \text{Var}(x) = \frac{1}{n}\sum_i (x_i - \bar{x})^2$$

Problem :- If x_i/f_i, $i=1(1)n$, is a frequency distribution of a variable x, then show that

$$\text{s.d.} = \frac{\sum_{i=1}^{n} f_i (k - x_i)^2}{\sum_{i=1}^{n} f_i} - s^2 \quad \text{where } \bar{x} = k + s,$$

and k is any constant.

Soln. →

$$s^2 = \frac{1}{\sum_{i=1}^{n} f_i} \sum_{i=1}^{n} (x_i - \bar{x})^2 f_i$$

$$= \frac{1}{\sum_{i=1}^{n} f_i} \sum_{i=1}^{n} \{k - x_i - (k - \bar{x})\}^2 f_i$$

$$= \frac{1}{\sum_{i=1}^{n} f_i} \sum_{i=1}^{n} (k - x_i)^2 f_i - (k - \bar{x})^2$$

$$= \frac{\sum_{i=1}^{n} f_i (k - x_i)^2}{\sum_{i=1}^{n} f_i} - s^2.$$

4> C.U

Explain standard deviation's superiority over other measures of dispersion. →

Ans:→ There are four absolute measures of dispersion – Range, SD, MD and QD. Among these SD is called superior because it possesses almost all the requisites of a good measure of dispersion.

1> It is rigidly defined. SD of a set of obsn. is the square root of the mean of square deviation from mean.

2> It is based on all obsn. Even if one of the obsn. is changed SD changes. However Range and QD don't possess this property.

3> Although not so simple as Range or QD, the calculation of SD is not very difficult and does not necessitate any special technique for computation. A change of origin does not effect SD.

4> SD is least affected by sampling flactuation. If several independent samples are drawn from some statistical population and each time all the four absolute measures of dispersion calculated, it will be found that SD fl uctuates less from sample to sample than any other measures of dispersion.

5> The unique property which makes SD superior to other measures is that it is amenable to algebric treatment.

Because of these advantages SD is almost exclusively used unless there are definite reasons for using other measures.

4> **Some Properties of the standard deviation :**–

i> If all the values of a variable are equal, its s.d is zero. The converse is also true. 5.4.

Let the variable x assume n values x_i, $i=1(1)n$ and let $x_i = c$, for each i.

$$\therefore \bar{x} = \frac{1}{n}\sum_{i=1}^{n} c = c.$$

Then $x_i - \bar{x} = 0$, for each i,

and $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} = \sqrt{\frac{1}{n}\cdot 0} = 0$.

On the contrary, suppose

$s = 0$, or $s^2 = 0$ or $\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = 0$

or, $\sum_{i=1}^{n}(x_i-\bar{x})^2 = 0$,

which is possible only when $x_i - \bar{x} = 0 \ \forall i$,

i.e. when $x_i = \bar{x} \ \forall i$.

Thus, if the s.d. is zero, all the values of the variable are equal.

ii) If $y = a + bx$ is the relation between two variables x and y, then their respective s.d., denoted by s_x and s_y, are related as $s_y = |b| s_x$. 5.2

We have the relation between two variables x and y as

$y = a + bx$

Hence $y_i = a + bx_i$, $i = 1(1)n$.

Then $\bar{y} = a + \bar{x}b$

So, $y_i - \bar{y} = b(x_i - \bar{x})$, $\forall i$

& $\sum_{i=1}^{n}(y_i-\bar{y})^2 = b^2 \sum_{i=1}^{n}(x_i-\bar{x})^2$

or, $\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 = b^2 \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$

or, $s^2_y = b^2 s_x^2$

Hence, $s_y = |b| s_x$.

From this property, it follows that the standard deviation is independent of change of origin (here a), but depends on change of scale (Here b).

iii) suppose two groups of values of a variable x are given. If $\bar{x}_1$ and s_1 respectively denote the mean and the s.d. of n_1 values contained in one group; $\bar{x}_2$ and s_2, the mean and s.d of n_2 values of the other group, then the standard deviation (s) of the combined data is given by

$$\boxed{s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1+n_2} + \frac{n_1(\bar{x}_1-\bar{x})^2 + n_2(\bar{x}_2-\bar{x})^2}{n_1+n_2}}$$

where the pooled mean $\bar{x} = \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1+n_2}$.

Suppose the values in the first group are $x_{1i}, i=1(1)n_1$, and those in the second group are $x_{2i}, i=1(1)n_2$. then

$$\bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i}, \quad \bar{x}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2i} \text{ and } \bar{x} = \frac{n_1\bar{x}_1+n_2\bar{x}_2}{n_1+n_2}.$$

Again, $s_1^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2$, $s_2^2 = \frac{1}{n_2}\sum_{i=1}^{n_2}(x_{2i}-\bar{x}_2)^2$.

The sum of squares of the deviations of the values of the two groups from $\bar{x}$ is

$$\sum_{i=1}^{n_1}(x_{1i}-\bar{x})^2 + \sum_{i=1}^{n_2}(x_{2i}-\bar{x})^2$$

Now, $$\sum_{i=1}^{n_1}(x_{1i}-\bar{x})^2 = \sum_{i=1}^{n_1}\{(x_{1i}-\bar{x}_1)+(\bar{x}_1-\bar{x})\}^2$$

$$= \sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2 + 2(\bar{x}_1-\bar{x})\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1) + n_1(\bar{x}_1-\bar{x})^2$$

$$= n_1 s_1^2 + n_1(\bar{x}_1-\bar{x})^2, \quad \text{since} \sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1) = 0.$$

Similarly, $\sum_{i=1}^{n_2}(x_{2i}-\bar{x})^2 = n_2 s_2^2 + n_2(\bar{x}_2-\bar{x})^2$.

Hence, we have,

$$s^2 = \frac{1}{n_1+n_2}\left\{\sum_{i=1}^{n_1}(x_{1i}-\bar{x})^2 + \sum_{i=1}^{n_2}(x_{2i}-\bar{x})^2\right\}$$

i.e. $$s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1+n_2} + \frac{n_1(\bar{x}_1-\bar{x})^2 + n_2(\bar{x}_2-\bar{x})^2}{n_1+n_2}. \quad \text{—— ①}$$

Putting $\bar{x} = \dfrac{n_1\bar{x}_1+n_2\bar{x}_2}{n_1+n_2}$ in ①, we get

$$s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1+n_2} + \frac{n_1 n_2(\bar{x}_1-\bar{x}_2)^2}{(n_1+n_2)^2}.$$

⇨ Quantiles and Percentiles :-

□ The quantile or fractile of order p (or the p-quantile) is a value of the variable such that a proportion p of the total number of given values are less than or equal to it and a proportion (1-p) are greater than or equal to it.
For continuous variable, this quantile (denoted by Z_p) may be approuimately determined by the formula,

$$Z_p = x_l + \frac{np - n_l}{f_0} \times c$$

where, x_l = lower boundary of the class-interval in which Z_p lies,

c = width of this class-interval,

n_l = cumulative frequency (of the 'less than-type) corresponding to x_l and

f_0 = frequency of this class.

□ The median is $Z_{1/2}$. The three values $Q_1 = Z_{1/4}$, $Q_2 = Z_{1/2}$, $Q_3 = Z_{3/4}$ which divide the frequency distribution of the variable into four equal parts are known as quartiles. Q_1, Q_2 and Q_3 are said to be the first (lower) quartile, second quartile, third (upper) quartile respectively.

□ Thereare nine deciles, denoted by $D_1, D_2, \ldots, D_9$, where $D_1 = Z_{1/10}$, $D_2 = Z_{2/10}$, ..., $D_9 = Z_{9/10}$. The deciles together devide the whole distribution into ten equal parts.

□ $P_1 = Z_{1/100}$, $P_2 = Z_{2/100}$, $P_3 = Z_{3/100}$, ..., $P_{99} = Z_{99/100}$ are 99 percentiles which jointly devide the frequency distribution of the variable into hundred equal parts.

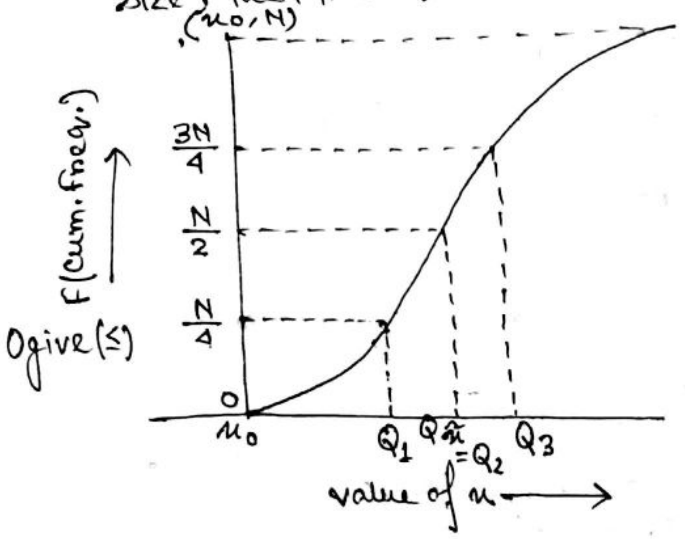
⇨ What is quartile deviation ?

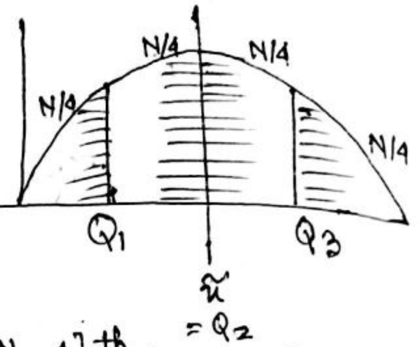
Ans:→ Quartile deviation is a measure of dispersion based on the quartiles. If the values of a variable differ much from one another, the differences between the quartiles would be large; on the other hand, when the values are close to one another, the differences would be small. As such one can take the average of the difference between Q_2 and Q_1 and that between Q_3 and Q_2 as a measure of dispersion, called quartile deviation (Q).

Thus, $Q = \frac{(Q_2 - Q_1) + (Q_3 - Q_2)}{2} = \frac{Q_3 - Q_1}{2}$,

It is also called the semi-interquartile range.

If the data are given in a frequency table with one or both of the terminal classes open or with class intervals of unequal size, then the quartile deviation is used as an appropriate measure.





$Q_2 = \left[\frac{N}{2}+1\right]^{th}$ obsn. $= x_{\left(\frac{N}{2}+1\right)}$

$Q_1 = \left[\frac{N}{4}+1\right]^{th}$ obsn. $= x_{\left(\frac{N}{4}+1\right)}$

$Q_3 = \left[\frac{3N}{4}+1\right]^{th}$ obsn. $= x_{\left(\frac{3N}{4}+1\right)}$

$\therefore Q_2 = x_l^{\textcircled{2}} + \frac{N/2 - CF^{\textcircled{2}}}{f_0^{\textcircled{2}}} \times c$ → corresponding to 2nd quartile class.

$\therefore Q_1 = x_l^{\textcircled{1}} + \frac{N/4 - CF^{\textcircled{1}}}{f_0^{\textcircled{1}}} \times c$ → corresponding to 1st quartile class.

$\therefore Q_3 = x_l^{\textcircled{3}} + \frac{3N/4 - CF^{\textcircled{3}}}{f_0^{\textcircled{3}}} \times c$ → corresponding to 3rd Quartile class.

$\therefore QD = \frac{Q_3 - Q_1}{2}$ = Quartile Deviation.

Property of Quartile Deviation :—

If two values x and y are related as $y = a + bx$, then —

$$Q_y = |b| Q_x .$$

When $b > 0$, $Q_3(y) = a + b.Q_3(x)$,

$Q_1(y) = a + b.\ Q_1(x)$.

$\therefore Q_3(y) - Q_1(y) = b\left[Q_3(x) - Q_1(x)\right]$

$\therefore \frac{Q_3(y) - Q_1(y)}{2} = b.\frac{Q_3(x) - Q_1(x)}{2}$, i.e. $Q(y) = b.Q(x)$. —— ①

Again, when $b < 0$, $Q_3(y) = a + b.Q_1(x)$,

and $Q_1(y) = a + b.Q_3(x)$.

$\therefore Q_3(y) - Q_1(y) = -b\{Q_3(x) - Q_1(x)\}$

or $\frac{Q_3(y) - Q_1(y)}{2} = -b.\frac{Q_3(x) - Q_1(x)}{2}$, i.e. $Q(y) = -b.Q(x$ —— ②

Combining ① and ②, we get —

$$Q(y) = |b|.Q(x) .$$

⇒ Empirical relation between mean and standard deviations :

Ans:→ For symmetrical or moderately skew distributions the mean deviation is about four fifths of the standard deviation.

⇒ Empirical relation between quartile and standard deviations :

Ans:→ For symmetrical and moderately skew distributions the quartile deviation is usually about two-thirds of the standard deviation.

Ques:→ Comparision of the measures of Dispersion :→

Ans:→ All the four measures of dispersion are rigidly defined. But the range becomes meaningless when one or both of the two limits of the variable values are infinite, and the quartile deviation may not be uniquely obtained in the case of ungrouped data or for the frequency distribution of a discrete variable.

The range is easy to compute. The other measures involve almost the same amount of computational labour. The significance of range, mean deviation and quartile deviation is easily comprehensible. But the standard deviation has a comparatively abstract nature.

Both the mean deviation and the standard deviation are based on all the obsn. of the variable. they characterise the whole set of values. the quartile deviation may remain unchanged even after the alternation of several values, the range is based on only on the two extreme values of the set.

The standard deviation has a number of discrete properties by virtue of which it is readily amenable to algebric treatment. But the other measures do not posses such properties.

In general, the standard deviation is the best measure of dispersion. Of course, range is preferred when speed of computation is of prime importance (as in the case of Statistical Quality control). Again, quartile deviation is the suitable measure in case of frequency distribution with open end classes.

Problems:

1) Prove that for any set of values $x_1, x_2, \ldots, x_n$, $\sum_{i=1}^{n} x_i^2 \geq \frac{(\sum x_i)^2}{n}$.

Solution:- C-S inequality: $\left(\sum_{i=1}^{n} a_i b_i\right)^2 \leq \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right)$

Let $a_i = x_i$, $b_i = 1$.

So, $$\left(\sum_{i=1}^{n} x_i\right)^2 \leq \left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} 1\right)$$

$$\Rightarrow \sum_{i=1}^{n} x_i^2 \geq \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}.$$

'=' holds when x_i's are all equal, $x_i = c$, $\bar{x} = c$.

2) Show that the standard deviation can't be smaller than the mean deviation about mean.

Solution:- Applying C-S inequality: $a_i = |x_i - \bar{x}|$ and $b_i = 1$.

$$\left\{\sum_{i=1}^{n} (x_i - \bar{x})^2\right\} \cdot n \geq \left\{\sum_{i=1}^{n} |x_i - \bar{x}|\right\}^2$$

$$\Rightarrow \left\{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2\right\} \geq \left\{\frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|\right\}^2$$; taking +ve sq. root, we have

$$\Rightarrow SD_x \geq MD_{\bar{x}}.$$

'=' holds when $|x_i - \bar{x}| = k \ \forall i$, i.e., $x_i = \bar{x} \pm k \ \forall \ i = 1(1)n$.

3) The difference between the AM and the median can't be greater than the standard deviation. or $|\bar{x} - Me| \leq s$.

Solution:-
$$\begin{aligned} |\bar{x} - Me| = \left|\frac{1}{n}\sum_{i=1}^{n} x_i - Me\right| &= \left|\frac{1}{n}\sum_{i=1}^{n}(x_i - Me)\right| \\ &= \frac{1}{n}\left|\sum_{i=1}^{n}(x_i - Me)\right| \\ &\leq \frac{1}{n}\sum_{i=1}^{n}|x_i - Me| \\ &\leq \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| \quad \text{(MD is least when measured about median)} \\ &\leq s. \end{aligned}$$

Implication of this result:- A measure without boundary can't be used for comparison of two or more freq. distn.. This result provides boundary to the measure.

4) Let x be a variable assuming the values $i=1,2,\dots,k$ with frequencies f_i and let F_i' be the corresponding cumulative frequencies of the "greater-than" type, while F_i'' are the cumulative totals of the "greater than" type of these cumulative frequencies. If n be the total frequency and $T_1 = \frac{1}{n}\sum_{i=1}^{k} F_i'$, $T_2 = \frac{1}{n}\sum_{i=1}^{n} F_i''$, show that $s^2 = 2T_2 - T_1 - T_1^2$.

Solution:- $T_1 = \bar{x}$

$$nT_2 = F_k'' + F_{k-1}'' + \dots + F_2'' + F_1'' = 1.F_1' + 2F_2' + \dots + (k-1)F_{k-1}' + kF_k'$$

$$= 1(f_1+f_2+\dots+f_k) + 2(f_2+f_3+\dots+f_k) + \dots + (k-1)(f_{k-1}+f_k) + kf_k$$

$$= 1.f_1 + (1+2)f_2 + (1+2+3)f_3 + \dots + (1+2+\dots+k)f_k$$

$$= \sum_{j=1}^{k}(1+2+\dots+j)f_j = \sum_{j=1}^{k}\frac{(1+j)j}{2}.f_j$$

$$\Rightarrow 2nT_2 = \sum_j jf_j + \sum_j j^2 f_j = nT_1 + \sum_j j^2 f_j$$

$$ns^2 = \sum j^2 f_j - n\bar{x}^2 = 2nT_2 - nT_1 - nT_1^2$$

$$\Rightarrow s^2 = 2T_2 - T_1 - T_1^2.$$

5) S.T. standard deviation is the least root mean square deviation.

Solution:- Let $f(A) = \sum_{i=1}^{n}(x_i - A)^2$

$$f'(A) = \sum_{i=1}^{n} 2(x_i - A)(-1)$$

$$f''(A) = -2\sum_{i=1}^{n}(-1) = 2n.$$

$$f'(A) = 0 \Rightarrow \sum_{i=1}^{n}(x_i - A) = 0 \Rightarrow \bar{x} = A \text{ and } f''(\bar{x}) > 0.$$

$$\therefore f(A) = \sum_{i=1}^{n}(x_i - A)^2 \text{ is minimum when } A = \bar{x}.$$

[For a given data $x_1, \dots, x_n$, $\frac{1}{n}\sum_{i=1}^{n}(x_i-A)^2$ depends on the choice of the average A, will give diffent values for different A. But the least value of $\frac{1}{n}\sum_{i=1}^{n}(x_i-A)^2$ is $\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ which is fixed for a given data $x_1, \dots, x_n$. Hence we define the standard deviation $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$ as a measure of dispersion.]

6) Let s and R be, respectively, the standard deviation and range of a set of n values of x. Show that

C.U

$$\frac{R^2}{2n} \le s^2 \le \frac{R^2}{4}.$$

5.10 (a)

When do the equalities hold?

Soln. →

i) Let the variable x assume the values $x_1, \dots, x_n$. Let the greatest and the least of the values be denoted by b and a. respectively.

Then $R = b - a$.

We know

$\sum_{i=1}^{n}(x_i-c)^2$ is least when $c = \bar{x}$.

Hence, $$\sum_{i=1}^{n}(x_i-\bar{x})^2 \le \sum\left(x_i - \frac{a+b}{2}\right)^2 = \sum\nolimits_1\left(x_i-\frac{a+b}{2}\right)^2 + \sum\nolimits_2\left(x_i-\frac{a+b}{2}\right)^2.$$

where Σ_1 and Σ_2 include respectively those values of x which are less than or equal to $\frac{a+b}{2}$ and greater than $\frac{a+b}{2}$.

$$\text{or } \sum_{i=1}^{n}(x_i-\bar{x})^2 \le \sum\nolimits_1\left(a-\frac{a+b}{2}\right)^2 + \sum\nolimits_2\left(b-\frac{a+b}{2}\right)^2$$

$$= \sum\nolimits_1 \frac{R^2}{4} + \sum\nolimits_2 \frac{R^2}{4} = n.\frac{R^2}{4}$$

Hence, $\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \le \frac{R^2}{4}$

i.e., $s^2 \le \frac{R^2}{4}$.

Here the equality sign holds either when all the values are equal or when the variable takes only two distinct values with same frequency.

[⊛ i) has been solved in 3. ii).]

Equality Case for $s^2 \leq \frac{R^2}{4}$:

Consider the frequency distribution :

Values	Freq
a	f_1
b	f_2
Total	n

$$\text{Mean} = \frac{af_1 + bf_2}{n}$$

$$ns^2 = \left(a - \frac{af_1+bf_2}{n}\right)^2 f_1 + \left(b - \frac{af_1+bf_2}{n}\right)^2 f_2$$

$$= \frac{(a-b)^2 f_2^2 f_1}{n^2} + \frac{(b-a)^2 f_1^2 f_2}{n^2}$$

$$= \frac{(a-b)^2 f_1 f_2}{n^2}\{f_1+f_2\}$$

$$= \frac{(a-b)^2 f_1 f_2}{n}$$

$$s^2 = \frac{(a-b)^2 f_1 f_2}{n^2} = \frac{(b-a)^2}{4}$$

$$\therefore n^2 = 4f_1f_2$$

$$\Rightarrow (f_1+f_2)^2 = 4f_1f_2$$

$$\Rightarrow f_1 = f_2$$

Hence $s^2 = \frac{R^2}{4}$ iff values x_i take only two values 'a' and 'b' with equal frequency.

ii) $$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

where $a \leq x_1 < x_2 < \dots\dots < x_n \leq b$

$$\geq \frac{1}{n}\{(x_1-\bar{x})^2 + (x_n - \bar{x})^2\}$$

$$\geq \frac{1}{n}\{(a-\bar{x})^2 + (b-\bar{x})^2\}$$

$$\left[(u^2+v^2) = \frac{1}{2}(u+v)^2 + \frac{1}{2}(u-v)^2 \geq \frac{1}{2}(u-v)^2 \right]$$

$$s^2 \geq \frac{1}{n}\{(a-\bar{x})^2+(b-\bar{x})^2\}$$

$$\geq \frac{1}{2n}\{(a-\bar{x})-(b-\bar{x})\}^2$$

$$= \frac{(b-a)^2}{2n}$$

$$\therefore s^2 \geq \frac{R^2}{2n}$$

The equality sign holds either if all the values of the variable are equal or if all the values except the maximum and minimum values are equal to $\frac{a+b}{2}$.

Hence, we have, $\boxed{\frac{R^2}{2n} \leq s^2 \leq \frac{R^2}{4}}$.

Equality Case for $\frac{R^2}{2n} \leq s^2$:

Consider the frequency distribution:

Values	freq
a	1
$\frac{a+b}{2}$	$n-2$
b	1
Total	n

$$\text{Mean} = \frac{a+\frac{a+b}{2}\cdot(n-2)+b}{n} = \frac{a+b}{2}$$

$$s^2 = \frac{1}{n}\left\{\left(a-\frac{a+b}{2}\right)^2\cdot 1 + \left(\frac{a+b}{2}-\frac{a+b}{2}\right)^2\cdot(n-2) + \left(b-\frac{a+b}{2}\right)^2\cdot 1\right\}$$

$$= \frac{(b-a)^2}{2n} = \frac{R^2}{2n}.$$

Q> Suppose that the variable x takes positive values only and that the deviations $x_i - \bar{x}$ are small compared to $\bar{x}$. Show that in such a case,

C.U

(a) $x_g \simeq \bar{x}\left(1 - \frac{s^2}{2\bar{x}^2}\right)$ and (b) $x_h \simeq \bar{x}\left(1 - \frac{s^2}{\bar{x}^2}\right)$ 5.15

$\boxed{**}$

Soln. →

(a) $$x_g = \left(\prod_{i=1}^{n} x_i\right)^{1/n}$$

$$\log x_g = \frac{1}{n}\sum_{i=1}^{n} \log x_i$$

$$= \frac{1}{n}\sum_{i=1}^{n} \log(x_i - \bar{x} + \bar{x})$$

$$= \frac{1}{n}\sum_{i=1}^{n} \log\left\{\bar{x}\left(1 + \frac{x_i - \bar{x}}{\bar{x}}\right)\right\}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left\{\log \bar{x} + \log\left(1 + \frac{x_i - \bar{x}}{\bar{x}}\right)\right\}$$

$$= \log \bar{x} + \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \frac{x_i - \bar{x}}{\bar{x}}\right)$$

$$= \log \bar{x} + \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{x_i - \bar{x}}{\bar{x}} - \frac{1}{2}\left(\frac{x_i - \bar{x}}{\bar{x}}\right)^2 + \cdots\cdots\right\}$$

[Since, $\frac{x_i - \bar{x}}{\bar{x}} \simeq 0$, i.e. $\left|\frac{x_i - \bar{x}}{\bar{x}}\right| < 1$]

$$\simeq \log \bar{x} + \left\{0 - \frac{1}{2\bar{x}^2}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}$$

[As $\frac{x_i - \bar{x}}{\bar{x}} \simeq 0$, neglecting $\left(\frac{x_i - \bar{x}}{\bar{x}}\right)^r$, $r \geq 3$.
Here $\sum_i (x_i - \bar{x}) = 0$

$$= \log \bar{x} - \frac{1}{2\bar{x}^2} \cdot s^2$$

$$= \log\left(\bar{x} \cdot e^{-s^2/2\bar{x}^2}\right)$$

$$\therefore x_g = \bar{x} \cdot e^{-s^2/2\bar{x}^2}$$

$$\simeq \bar{x}\left\{1 - \frac{s^2}{2\bar{x}^2}\right\}$$

[$e^{-x} \simeq 1 - x$ for small x.
$\frac{s^2}{2\bar{x}^2} = \frac{1}{2} \cdot \frac{1}{n}\sum\left(\frac{x_i - \bar{x}}{\bar{x}}\right)^2$ are small]

(b) $$x_h = \frac{n}{\sum\limits_{i=1}^{n}\frac{1}{x_i}}$$

$$= \frac{n}{\sum\limits_{i=1}^{n}\frac{1}{x_i-\bar{x}+\bar{x}}}$$

$$= \frac{n}{\sum\limits_{i=1}^{n}\frac{1}{\bar{x}\left\{1+\frac{x_i-\bar{x}}{\bar{x}}\right\}}}$$

$$= \frac{n\bar{x}}{\sum\limits_{i=1}^{n}\left(1+\frac{x_i-\bar{x}}{\bar{x}}\right)^{-1}}$$

$$= \frac{n\bar{x}}{\sum\limits_{i=1}^{n}\left\{1-\left(\frac{x_i-\bar{x}}{\bar{x}}\right)+\left(\frac{x_i-\bar{x}}{\bar{x}}\right)^2 - \cdots\cdots\right\}}$$ [since $\frac{x_i-\bar{x}}{\bar{x}} \simeq 0 \;\forall i$, i.e. $\left|\frac{x_i-\bar{x}}{\bar{x}}\right| < 1$]

$$\simeq \frac{n\bar{x}}{\sum\limits_{i}\left\{1-\left(\frac{x_i-\bar{x}}{\bar{x}}\right)+\left(\frac{x_i-\bar{x}}{\bar{x}}\right)^2\right\}}$$ [As $\frac{x_i-\bar{x}}{\bar{x}}$ are small, neglecting $\left(\frac{x_i-\bar{x}}{\bar{x}}\right)^n, n \geq 3$]

$$= \frac{n\bar{x}}{n - 0 + \frac{\sum_i (x_i-\bar{x})^2}{\bar{x}^2}}$$

$$= \frac{\bar{x}}{1+\frac{s^2}{\bar{x}^2}}$$

$$= \bar{x}\left(1+\frac{s^2}{\bar{x}^2}\right)^{-1}$$

$$\simeq \bar{x}\left(1-\frac{s^2}{\bar{x}^2}\right)$$ [$(1+x)^{-1} \simeq 1-x$ for small x. Here, $\frac{s^2}{\bar{x}^2} = \frac{1}{n}\sum_i\left(\frac{x_i-\bar{x}}{\bar{x}}\right)^2$ is small]

Relative Measures of Dispersion :

Whenever we want to compare the variability of the two data sets which differs widely in their averages or which are measured in different units, we donot simply calculate the absolute measures of dispersion but we calculate the relative measures of dispersion which are independent of units of the variables.

Suppose repeated measurements are being taken of two rods, one of length 10cm and another of length 100 cm approximately. Suppose the standard deviation of each set of measurements be 20. Two standard deviations are equal, but the first set of measurements is much less accurate than the second set of measurements. Clearly, the quantity $\frac{\text{S.d}}{\text{Mean}} \times 100$ will give a true picture of their variability and this measure is known as Coefficient of variation. the C.V. is a pure no. independent of the units of measurements.

⟨i⟩ Coefficient of variation :

$$C.V = \frac{\text{Standard deviation}}{\text{Mean}} \times 100$$

⟨ii⟩ Co-efficient of mean deviation :

$$C.D_M = \frac{\text{Mean deviation about mean (or median)}}{\text{Mean(or median)}} \times 100$$

⟨iii⟩ Coefficient of quartile deviation :

$$C.D_Q = \frac{\text{Quartile deviation}}{\text{Median}} \times 100$$

Uses :→ It is used primarily for comparing dispersions of variables having different units of measurements. It is basically used for comparing the variability of two series which differ widely in their averages or which are measured in different units, we calculate the C.V's for each series. The data having greater C.V. is said to be more variable than the other and the data having lesser C.V. is said to be more consistent (or homogeneous) than the other.

— This measure is based on dispersion and central tendency.

C.U

Ques:→ Distinguish between absolute and relative measures of dispersion.

Ans:→ Measures of dispersion of two types – Absolute and Relative.

Absolute measures are – Range, QD, SD & MD.

Relative measures are – Coefficient of – i) variation,
ii) mean deviation,
iii) quartile deviation.

1) Absolute measures are expressed in the same unit in which the obsn are given. Relative measures are obtained by expressing an absolute measure as percentage of a measure of central tendency and hence Relative measures are independent of the units of measurements.

2) Usually the absolute measures are employed for measuring dispersion. But for purpose of comparing dispersion in different series, relative measures are used. Again when two sets of data are given in dissimilar units, there is no other alternative but to use a relative measures for comparison.

3) Relative measures may also be used to compare the relative accuracy of the data but Absolute measures can not be so used.

4) Measures based on mutual differences of observations:-

Since dispersion really means the extent to which the given values of the variable differ from one another (rather than from any arbitrarily chosen average), any proper measure of dispersion should be independent of measure of central tendency and should be based solely on the mutual differences $x_i - x_j$ $(i, j = 1, 2, \dots, n)$ of the values of x.

(a) A measure of this type, suggested by Gini, is the mean of the absolute values of all n^2 mutual differences. Called Gini's mean difference, given by the formula –

$$\Delta_1 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j| .$$

(b) A measure of similar type which is based on the squares of mutual differences of the obsn., is —

$$\Delta_2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i-x_j)^2$$

Note that, —

$$\therefore n^2 . \Delta_2 = \sum_{i=1}^{n}\sum_{j=1}^{n}(x_i-x_j)^2$$

$$= \sum_{i}\sum_{j}\{(x_i-\bar{x})-(x_j-\bar{x})\}^2$$

$$= \sum_{i}\sum_{j}\{(x_i-\bar{x})^2+(x_j-\bar{x})^2-2(x_i-\bar{x})(x_j-\bar{x})\}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}(x_i-\bar{x})^2+\sum_{i=1}^{n}\sum_{j=1}^{n}(x_j-\bar{x})^2-2\sum_{i=1}^{n}(x_i-\bar{x})\sum_{j=1}^{n}(x_j-\bar{x})$$

$$= \sum_{j=1}^{n}ns^2+\sum_{i=1}^{n}ns^2-2\times 0 \quad \left[\because \sum_{i=1}^{n}(x_i-\bar{x})=0\right]$$

$$= 2n^2s^2$$

$$\text{i.e. } \cancel{n^2}.\Delta_2 = 2\cancel{n^2}s^2$$

$$\Leftrightarrow \Delta_2 = 2s^2$$

$$\therefore s^2 = \frac{\Delta_2}{2}.$$

[It is the one more reason why the S.D. should generally be regarded as the best measure of dispersion]

$\boxed{\text{C.U}}$

Ques :→ Show that, if s^2 be the variance of n given values $x_1, x_2, \ldots, x_n$ of a variable x, then —

$$s^2 = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i-x_j)^2.$$

Hence or otherwise examine the consequence of additing a constant value to all the observations.

$\boxed{\text{Or}}$

Show that — $\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i-x_j)^2 = 2\,\text{Var}(x)$.

Sol. :- Try Yourself.

Lorenz Curve or Curve of Concentration

Ques:-

Explain the concept of Lorenz curve or Curve of Concentration, area of concentration and line of equal distribution. Discuss the use of Lorenz curve? [CU]

ANS:- The degree of inequality in the distribution of some variable x (income, wealth, etc.) or the concentration of the variable x is represented in terms of a special type of cumulative frequency curve, known as a Curve of Concentration or Lorenz curve.

Let the total frequency of the distribution be N and its total value (total income) be I

Define $F(x) = \left\{ \dfrac{\text{Number of values which are} \le x}{N} \right\} \times 100\%$, as the percentage of cumulative frequency for the variable upto the value x and

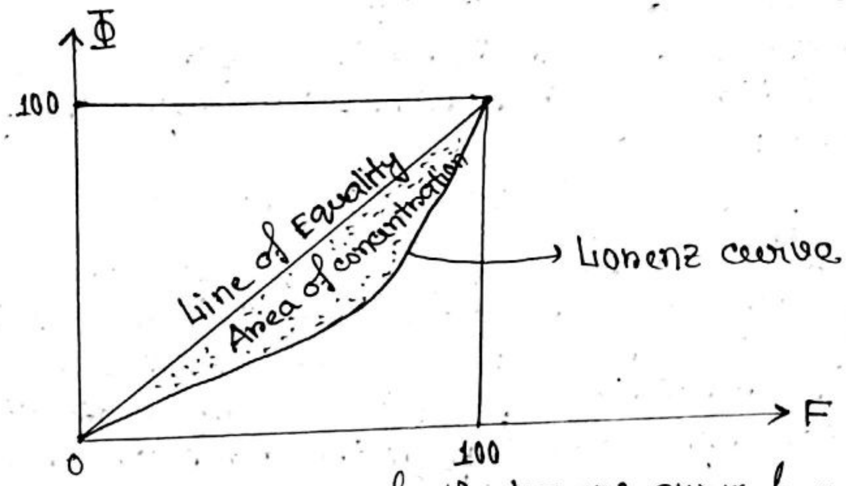
$$\Phi(x) = \left\{ \frac{\text{Total of the variable values which are} \le x}{I} \right\} \times 100\%,$$

as the percentage of cumulative total of the variable up to the value x.

Clearly, F and Φ vary from 0 to 100. The curve obtained by $\Phi(x)$ against $F(x)$ for different fixed values of x, is known as the Lorenz curve or Curve of Concentration. So, Lorenz curve shows the relationship between $F(x)$ and $\Phi(x)$.

The straight line $\Phi = F$ is known as the 'Line of equal distribution', since $\Phi(x) = F(x)$ implies that any particular proportion of members coould have the same proportion of total value. In the case of an income distribution, $\Phi(x) = F(x) = 20\%$, it coould mean that 20% of persons coould earn or possess 20% of the total income.

In case of unequal distribution lower income groups coould have proportionately lower share in the total income. Hence, in the beginning the slope of the Lorenz curve is smaller than the slope of the line of equality. As a result the Lorenz curve must lie below the diagonal line.



The greater the departure of the Lorenz curve from the Line of equal distn., the higher is the concentration of the variable(or income) values in a few members. In particular, if we find that 90% of the persons receive only 50% of the ~~total~~ income and consequently the remaining 10% of the persons possess 50% of the total income, it means there is a high degree of concentration in a few members in the upper income groups.

Hence the area between the line of equality and the curve of concentration or Lorenz curve is called the area of concentration, is a measure of the degree of concentration of the variable (income, wealth, etc.); the larger the area the more is the concentration.

Uses of Lorenz Curve:- Sometimes, we are interested to know how inequality in distribution has changed overtime or how inequalities compared as between two countries, etc. When one Lorenz curve dominates the other through out its range, the Lorenz curve which is uniformly higher, that is to say, nearer the line of equality compared to the other, implies less inequality.

In case two Lorenz curves intersect, this Lorenz curve domination criterion breaks down and then we can compare the areas of concentration of the curves. The distribution with the greater area of concentration will have less inequality in distribution among the variable values.

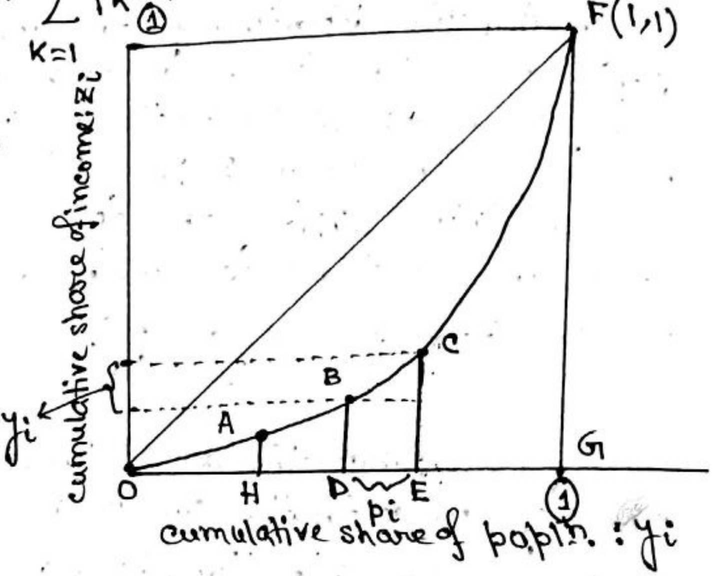
Computation of Area of Concentration:-

The statistical data on distribution of income and wealth is usually available as grouped data. Let us define the following new variables :

$$p_i = \frac{f_i}{N}, \; x_i = \frac{I_i}{I}, \; z_i = \sum_{k=1}^{i} x_k, \; i = 1, 2, \ldots, n.$$

where, N is the total number of persons, f_i is the number of persons in the ith income class; I is the total income and I_i is the income of the ith class, x_i be the corresponding income share and z_i is the cumulative share of income up to the ith income class. We also define $y_i = \sum_{k=1}^{i} p_k$ ① as the cumulative share of population for the ith class.

We develop a computational formula for the area of concentration from the empirical Lorenz curve.



①
F(1,1)
cumulative share of income: z_i
C
B
A
y_i
G
O
H
D
E
p_i
①
cumulative share of popln. : y_i

Hence, the area bounded by the Lorenz curve is given by :

$$= \frac{1}{2} p_1 z_1 + \sum_{i=2}^{n} p_i \left(\frac{z_i + z_{i-1}}{2} \right),$$

$$= \frac{1}{2} \sum_{i=1}^{n} p_i \left(z_i + z_{i-1} \right), \text{ where } z_0 = 0.$$

[Since, $\Delta AHO = \frac{1}{2} \times \text{base} \times \text{height} = \frac{1}{2} \times p_1 \times z_1$
and Area of trapizium BCED = $\frac{1}{2} \times$ (sum of the parallel sides) $\times$ (distance between the parallel sides)

$$= \frac{1}{2} \times (CE + BD) \times DE$$

$$= \frac{1}{2} (z_i + z_{i-1}) p_i \;]$$

Therefore, the area of concentration = {the area of the triangle OFG} − {the area bounded by the lorenz curve}

$$= \frac{1}{2} - \frac{1}{2} \sum_{i=1}^{n} p_i (z_{i-1} + z_i), \text{ where } z_0 = 0.$$

Gini's coefficient of Concentration:

Let $x_1, \ldots, x_n$ be n values of a variable x.

A measure suggested by Gini, is the mean of the absolute values of all n^2 mutual differences, called Gini's mean difference and it is given by

$$\Delta_1 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|$$

Gini's coefficient of concentration (G) is exactly $\frac{1}{2}$ of the relative mean differences, i.e. $\frac{\Delta_1}{\bar{x}}$.

$$\text{Thus,} \quad G = \frac{\Delta_1}{2\bar{x}} = \frac{1}{2n^2\bar{x}} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|.$$

This formula represents G as a weighted sum of values (on individual's income) where weights are the difference of their ranks from their median rank.

Problem:- If income of each person is increased by the same amount, then what will be the effect on Δ_1 and Gini-coefficient of concentration?

Solution:- Let $y_i = x_i + c$, $i = 1(1)n$.

$$\text{Then} \quad \Delta_1^y = \frac{1}{n^2} \sum\sum |y_i - y_j|$$
$$= \frac{1}{n^2} \sum\sum |x_i + c - x_j - c|$$
$$= \frac{1}{n^2} \sum\sum |x_i - x_j|$$
$$= \Delta_1^x.$$

$$\text{By} \quad G_y = \frac{\Delta_1^y}{2\bar{y}} = \frac{\Delta_1^x}{2(\bar{x}+c)} < \frac{\Delta_1^x}{2\bar{x}} = G_x, \text{ as } c > 0.$$

Gini's coefficient and Lorenz Curve : —

The statistical data on distribution of income and wealth is usually available as grouped data. therefore, for the purpose of statistical analysis it is necessary to define Gini's coefficient for grouped data. In this context we shall also explore the close relationship between Gini coefficient and Lorenz curve.

Define, $p_i = \frac{f_i}{N}$, $y_i = \sum_{k=1}^{i} p_k$, $x_i = \frac{I_i}{I}$, $z_i = \sum_{k=1}^{i} x_k$,

where, N is the total number of persons, f_i is the number of persons in the i^{th} income class ; I is the total income and I_i is the income in the i^{th} income class, x_i is the corresponding income share and z_i is the cumulative share of Income upto the i^{th} income class.

If income classes are arranged from the bottom of the distribution then it is clear that

$$\frac{x_1}{p_1} \le \frac{x_2}{p_2} \le \frac{x_3}{p_3} \le \cdots\cdots \le \frac{x_n}{p_n}, \text{ since } \frac{x_i}{p_i} = \frac{I_i/f_i}{I/N}.$$

is the ratio of the per-capita income of the i^{th} class to the per-capita income of the total population.

Then the Gini's coefficient of concentration is given by

$$G = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j \left| \frac{x_i}{p_i} - \frac{x_j}{p_j} \right|$$

Note that,

$$G = \frac{1}{2}\cdot 2\sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j \left(\frac{x_i}{p_i} - \frac{x_j}{p_j} \right)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} (x_i p_j - x_j p_i)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} x_i p_j - \sum_{i=1}^{n} p_i \sum_{j=1}^{i} x_j$$

$$= \sum_{j=1}^{n}\sum_{i=j}^{n} x_i p_j - \sum_{i=1}^{n} p_i z_i$$, interchanging the order of summations in the first sum, where, $1 \le j \le i \le n$.

$$= \left(\sum_{i=j}^{n} x_i\right)\sum_{j=1}^{n} p_j - \sum_{i=1}^{n} p_i z_i$$

$$= \sum_{j=1}^{n} p_j \left(1 - \sum_{i=1}^{j-1} x_i\right) - \sum_{i=1}^{n} p_i z_i$$, since $\sum_{j=1}^{n} p_j = 1$.

$$= 1 - \sum_{j=1}^{n} p_j z_{j-1} - \sum_{i=1}^{n} p_i z_i$$

$$\therefore G = 1 - \sum_{i=1}^{n} p_i \left(z_{i-1} + z_i\right), \quad z_0 = 0.$$

$$= 2 \times \{\text{the area of concentration}\}$$

Hence, the Gini's coefficient of concentration is twice the area of concentration.

In terms of Lorenz curve,

$$G = \frac{\text{the area between the lorenz curve \& line of equality}}{\text{the area of the triangle below the line of equality}},$$

since the area below the line of equality is $1/2$.

Hence, Gini coefficient is also known as Lorenz ratio or concentration ratio because of this relationship.

Problem:- The coefficient of variation of x is defined as $\frac{s}{\bar{x}}$. If $x_1, \ldots, x_n$ are the n values of x, show that the C.V. is the standard derivation of u_i's, where $u_i = \frac{x_i - \bar{x}}{\bar{x}}$.

Solution:-

$$u_i = \frac{x_i - \bar{x}}{\bar{x}}, \quad \bar{u} = \frac{1}{n} \sum u_i = \frac{1}{n\bar{x}} \sum (x_i - \bar{x}) = 0$$

Hence, $$s_u^2 = \frac{1}{n} \sum u_i^2 - \bar{u}^2 = \frac{1}{n} \sum \left(\frac{x_i - \bar{x}}{\bar{x}}\right)^2$$

$$= \frac{1}{\bar{x}^2} \cdot \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$= \frac{s^2}{\bar{x}^2}.$$

$$\Rightarrow s_u = \frac{s}{\bar{x}}.$$

CHAPTER - 4

Moments and Measures of Skewness and Kurtosis

C.U

MOMENTS :- 6.1.

1. Define raw moments of order r ?

Case I → Non-frequency type data :

Let $x_1, x_2, \ldots, x_n$ be n values of a variable x. Then the rth order raw moment of x about A is defined by —

$$m_r'(A) = \frac{1}{n}\sum_{i=1}^{n}(x_i - A)^r \quad \text{——— (i)}$$

In particular, when A = 0, then the rth order raw moment of x is defined by, —

$$m_r'(0) = \frac{1}{n}\sum_{i=1}^{n} x_i^r \quad \text{——— (ii)}$$

Usually, $m_r'(0)$ is simply denoted by m_r'.

Note that, $m_1' = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the mean of the distribution.

(*) Case II → Frequency type data :

Let $x_1, x_2, \ldots, x_k$ be k observations with frequencies $f_1, f_2 \ldots f_k$.

So, $\sum_{i=1}^{k} f_i = n$

∴ rth order raw moment about the origin A is defined by,

$$m_r'(A) = \frac{1}{n}\sum_{i=1}^{k}(x_i - A)^r f_i$$

In particular, $m_1' = \frac{1}{n}\sum_{i=1}^{k} x_i f_i$ is the mean of the distribution.

(*) 2. Define central moment of order r ?

When the origin of a moment is taken at the AM of the variable, it is called a central moment.

If $A = \bar{x}$, then $m_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r = m_r'(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_1')^r$ is known as the rth order central moment of x.

For grouped data, $m_r = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^r \cdot f_i$

Note that, The mean of a variable is its first order moment about 0 while the variance of the frequency distribution is the second order central moment. Also it's worth noting that → $m_0' = m_0 = 1$ and $m_1 = 0$.

C.U. ⇨ Central moments in terms of raw moments :—

Ques:→ (*) Express the rth order central moment in terms of the rth and lower order raw moments.

Ans:→ The rth order central moment is —

$$m_r = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^r$$

$$= \frac{1}{n}\sum_{i=1}^{n}\{(x_i-A)-(\bar{x}-A)\}^r$$

$$= \frac{1}{n}\sum_{i=1}^{n}\{(x_i-A)-m_1'(A)\}^r \qquad \left[\because m_1'(A)=\frac{1}{n}\sum_{i=1}^{n}(x_i-A) = \bar{x}-A\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\Big\{(x_i-A)^r - \binom{r}{1}(x_i-A)^{r-1}m_1'(A) + \binom{r}{2}(x_i-A)^{r-2}m_1'^2(A)$$
$$- - - - - - - - + (-1)^r\binom{r}{r}m_1'^r(A)\Big\}$$

$$= m_r'(A) - \binom{r}{1}m_{r-1}'(A)\cdot m_1'(A) + \binom{r}{2}m_{r-2}'(A)\cdot m_1'^2(A) - - - -$$
$$- - - - + (-1)^r m_1'^r(A).$$

$$\therefore \boxed{\begin{aligned} m_r &= m_r'(A) - \binom{r}{1}m_{r-1}'(A)m_1'(A) + \binom{r}{2}m_{r-2}'(A)\cdot m_1'^2(A) - - - - \\ &\quad - - - - + (-1)^r m_1'^r(A) \\ &= \sum_{k=0}^{r}(-1)^k\binom{r}{k}m_{r-k}'(A)\cdot[m_1'(A)]^k \quad \text{———— (1)} \end{aligned}}$$

If we take $A=0$, then —

$$m_r = \sum_{k=0}^{r}(-1)^k\binom{r}{k}m_{r-k}'(m_1')^k.$$

Some particular cases of (1) are:

$m_0 = 1$

$m_1 = m_1' - m_1' = 0$

$m_2 = m_2' - 2m_1'm_1' + m_1'^2 = m_2' - m_1'^2,$

$m_3 = m_3' - 3m_2'm_1' + 3m_1'm_1'^2 - m_1'^3 = m_3' - 3m_2'm_1' + 2m_1'^3,$

$m_4 = m_4' - 4m_3'm_1' + 6m_2'm_1'^2 - 4m_1'm_1'^3 + m_1'^4$

$\quad = m_4' - 4m_3'm_1' + 6m_2'm_1'^2 - 3m_1'^4$

⇨ Raw moments in terms of Central moment :—

□ 1st method :~ $m_r'(A) = \frac{1}{n}\sum_{i=1}^{n}(x_i - A)^r$

(*)

$$= \frac{1}{n}\sum_{i=1}^{n}\left[(x_i-\bar{x})+(\bar{x}-A)\right]^r$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[(x_i-\bar{x})+m_1'(A)\right]^r$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=0}^{r}\binom{r}{k}(x_i-\bar{x})^{r-k}\{m_1'(A)\}^k\right]$$

$$= \sum_{k=0}^{r}\binom{r}{k}.\, m_{r-k}\, .\, m_1'^{k}(A).$$

(or)

□ 2nd Method :~

$$m_r' = \frac{1}{n}\sum_{i=1}^{n}(x_i-A)^r$$

$$= \frac{1}{n}\sum_{i=1}^{n}\{(x_i-\bar{x})+(\bar{x}-A)\}^r$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[(x_i-\bar{x})+d\right]^r \quad \text{, where } d=\bar{x}-A.$$

$$= \frac{1}{n}\sum_{i=1}^{n}\Big\{(x_i-\bar{x})^r + \binom{r}{1}(x_i-\bar{x})^{r-1}d + \binom{r}{2}(x_i-\bar{x})^{r-2}d^2 + \cdots$$
$$\cdots\cdots + \binom{r}{r-1}(x_i-\bar{x})d^{r-1} + d^r\Big\}$$

$$= m_r + \binom{r}{1}m_{r-1}d + \binom{r}{2}m_{r-2}d^2 + \cdots\cdots\cdots\cdots$$
$$+ \binom{r}{r-2}m_2 d^{r-2} + d^r \quad \left[\because m_1=0\right]$$

———— ②

Some particular cases of (2) are :

$m_1'(A) = m_1 + d = d$

$m_2'(A) = m_2 + 2m_1 d + d^2 = m_2 + d^2$

$m_3'(A) = m_3 + 3m_2 d + 3m_1 d^2 + d^3 = m_3 + 3m_2 d + d^3.$

$m_4'(A) = m_4 + 4m_3 d + 6m_2 d^2 + 4m_1 d^3 + d^4$

$= m_4 + 4m_3 d + 6m_2 d^2 + d^4 \quad$ [Since, $m_0 = 1$, $m_1 = 0$]

Properties of moments :→

i) If $x_1, x_2, \ldots, x_n$ are n observation such that $x_i = c$ for all;
then $m_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r = 0 \ \forall r$.

(*) ii) **Effect of change of the origin and scale on central moments :→**

Let $x_1, x_2, \ldots, x_n$ be the n values of a variable x. Consider the change in the origin and scale of the variable x i.e. let $y = a + bx$.

Then, $y_i = a + bx_i \quad , i = 1(1)n$.

$$m_r(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^r$$

$$= \frac{1}{n}\sum_{i=1}^{n}(a + bx_i - a - b\bar{x})^r$$

$$= \frac{b^r}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r$$

i.e. $\boxed{m_r(y) = b^r \cdot m_r(x)}$

$\boxed{C.V}$

iii) $m_4 \geq m_2^2$.

Soln → To prove the above inequality let us consider the cauchy - schwarz inequality -

$$\left(\sum a_i^2\right)\left(\sum b_i^2\right) \geq \left(\sum a_i b_i\right)^2 \text{ ——— } (1)$$

Let us take, $a_i = (x_i - \bar{x})^2\sqrt{f_i} \quad , b_i = \sqrt{f_i}$

$\therefore \sum a_i^2 = \sum (x_i - \bar{x})^4 \cdot f_i \quad ; \sum b_i^2 = \sum f_i = N$ (say)

And $\sum a_i b_i = \sum (x_i - \bar{x})^2 \cdot f_i$.

From (1), we get — $\sum (x_i - \bar{x})^4 f_i \cdot N \geq \left[\sum (x_i - \bar{x})^2 f_i\right]^2$

$$\Rightarrow N^2 \cdot m_4 \geq N^2 \cdot m_2^2$$

i.e. $m_4 \geq m_2^2$ (Proved)

□ **Equality ('=') holds when —**

$$\frac{a_i}{b_i} = k \text{ (constant)}$$

i.e. $$\frac{(x_i - \bar{x})^2 \sqrt{f_i}}{\sqrt{f_i}} = k$$

$$\Rightarrow x_i = \bar{x} \pm c$$

Note that x_i takes any two values $\bar{x}+c$ and $\bar{x}-c$. For $\bar{x}$ being the mean of $x_1, x_2, \dots, x_n$, we must have equal frequency for two values.

because, $x_i = \begin{cases} \bar{x}-c & ; f_1 \\ \bar{x}+c & ; f_2 \end{cases}$

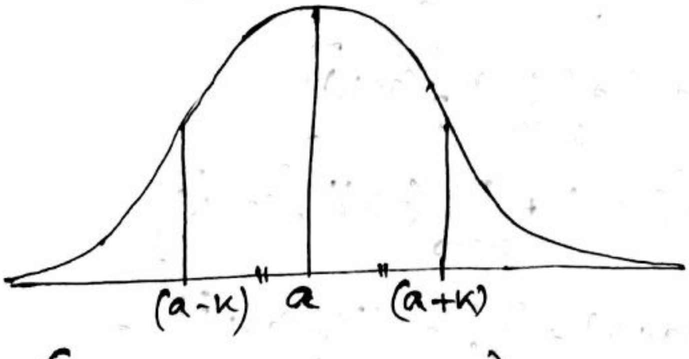
$$\text{Mean} = \frac{(\bar{x}-c)f_1 + (\bar{x}+c)f_2}{f_1+f_2}$$

$$= \frac{\bar{x}(f_1+f_2) - c(f_1-f_2)}{(f_1+f_2)}$$

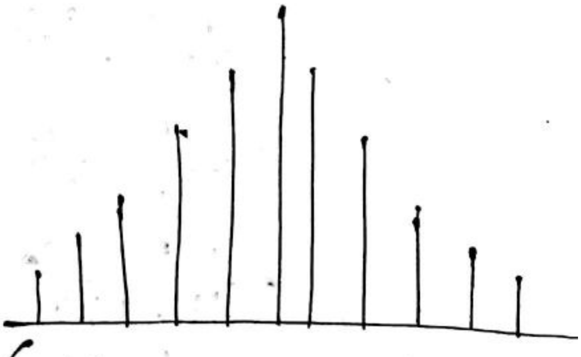
$$= \bar{x} - c.\frac{f_1-f_2}{f_1+f_2}.$$

Symmetrical Frequency Distribution :-

For a distribution symmetric about $x=a$, we must have $f(a+k) = f(a-k)$, ∀ $k>0$, where $f(x)$ stand for the frequency of $X=x$.



(If x is continuous)



(If x is discrete)

(*) 5) **Result :→ For a symmetric distribution every odd order central moments are zero.**

or,

All odd ordered central moments of a symmetrical frequency distribution are zero.

(*) Proof :→ Let the distribution be symmetric about $x = a$ i.e. $f(a+k) = f(a-k)\ \forall\ k$, where $f(x)$ stands for the frequency $x = a$.

$$\text{Mean}(\bar{x}) = \frac{1}{N}\sum_{k}(a+k)f(a+k) \quad , \text{ where } N = \sum_{k} f(a+k)$$

$$= \frac{1}{N}\Big[\sum_{k<0}(a+k)f(a+k) + \sum_{k>0}(a+k)f(a+k) + a.f(a)\Big]$$

$$= \frac{1}{N}\Big[a\underbrace{\Big\{\sum_{k<0}f(a+k) + \sum_{k>0}f(a+k) + f(a)\Big\}}_{\left[=\sum f(a+k) = N\right]} + \sum_{k<0}k.f(a+k) + \sum_{k>0}k.f(a+k)\Big]$$

$$= \frac{1}{N}\Big[a\Big\{\sum_{k}f(a+k)\Big\} + \sum_{k<0}k.f(a+k) + \sum_{k>0}k.f(a+k)\Big]$$

$$= \frac{1}{N}\Big[aN + \underbrace{\sum_{k>0}(-k)f(a-k) + \sum_{k>0}k.f(a+k)}_{=0}\Big] \quad \left[\text{as } f(a+k) = f(a-k) \text{ for symmetric distribution}\right]$$

$$= \frac{1}{N}(aN)$$

$$\text{i.e. } \bar{x} = a. \quad = a$$

$$m_r = \frac{1}{N}\sum_{k}(a+k-\bar{x})^r.f(a+k)$$

$$= \frac{1}{N}\sum_{k}(\cancel{a}+k-\cancel{a})^r f(a+k) \quad [\because \bar{x} = a]$$

$$= \frac{1}{N}\Big[\sum_{k>0}k^r f(a+k) + \sum_{k<0}k^r f(a+k)\Big]$$

$$= \frac{1}{N}\Big[\sum_{k>0}(k)^r f(a+k) + \sum_{k>0}(-k)^r f(a-k)\Big]$$

$$= \frac{1}{N}\Big[\sum_{k>0}(k)^r f(a+k) + \sum_{k>0}(-k)^r f(a+k)\Big] \quad [\because f(a+k) = f(a-k)]$$

$$= \frac{1}{N}\sum_{k>0}\big[(+k)^r + (-k)^r\big]f(a+k)$$

when r = odd, then $m_r = 0$.

Theorem :→ For a unimodal freq. distn. the mean, mode and median coincide.

Proof :→ Consider a unimodal symmetric freq distn. about x_0 :

Values	Freq.
x_0-h_n	f_n
$\vdots$	$\vdots$
x_0-h_1	f_1
x_0	f_0
x_0+h_1	f_1
$\vdots$	$\vdots$
x_0+h_n	f_n
Total=	N

$$m_1'(x_0) = \frac{1}{N}\sum_i (x_i - x_0) f_i$$

$$= \frac{1}{N}\{(x_0 - h_n - x_0) f_n + \cdots\cdots + (x_0 - h_1 - x_0) f_1 + (x_0 - x_0) f_0 + \cdots\cdots + (x_0 + h_n - x_0) f_n\}$$

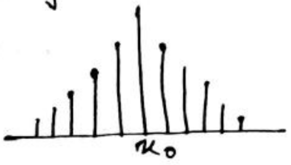
$$= \frac{1}{N}\{(-h_n f_n) + \cdots\cdots + (-h_1 f_1) + 0 + h_1 f_1 + \cdots\cdots + h_n f_n\}$$

$$= 0 .$$

$$\Rightarrow \frac{1}{N}\sum x_i f_i = x_0$$

$$\Rightarrow \bar{x} = x_0$$

A unimodal freq. distn. is of the form:



The value x_0 has the maximum frequency, from graph,
Hence, mode of x is x_0.
Note, {the no. of values x_i which are $\leq x_0$}

$$= \{f_n + f_{n-1} + \cdots\cdots + f_1 + f_0\}$$

$$= \{\text{the no. of values } x_i \geq x_0\}.$$

By defn. , x_0 is the median of x.

Therefore, for a unimodal symmetric freq. distn. ,

mean = median = mode
= the point of symmetry (x_0).

C.U

Sheppard's Correction for moments / Corrections for grouping :

When moments are calculated from a numerically specified distribution which is grouped, there is present a certain amount of approximation owing to the fact that the frequencies are assumed to be concentrated at the mid-points of intervals. In practice the frequencies may not be concentrated at the mid points, the assumption naturally introduces some errors which are called the error due to grouping.

To correct for these grouping errors, the computed values of the moments have to be suitably adjusted. A method for adjusting the moments for grouped data where the classes are equally wide has been developed by Sheppard.

Sheppard's corrections for moments about an arbitrary origin are given below:

$$m_1' \text{(corrected)} = m_1'$$
$$m_2' \text{(corrected)} = m_2' - \frac{c^2}{12},$$
$$m_3' \text{(corrected)} = m_3' - \frac{c^2}{4} m_1'.$$
$$m_4' \text{(corrected)} = m_4' - \frac{c^2}{2} m_2' + \frac{7}{240} c^4.$$

and for central moments,

$$m_2 \text{(corrected)} = m_2 - \frac{c^2}{12},$$
$$m_3 \text{(corrected)} = m_3,$$
$$m_4 \text{(corrected)} = m_4 - \frac{c^2}{2} m_2 + \frac{7}{240} c^4,$$

where c is the width of each class-interval.

C.U These corrections will be valid if certain conditions are fulfilled :

i) It is necessary that the observations should relate to continuous variable.

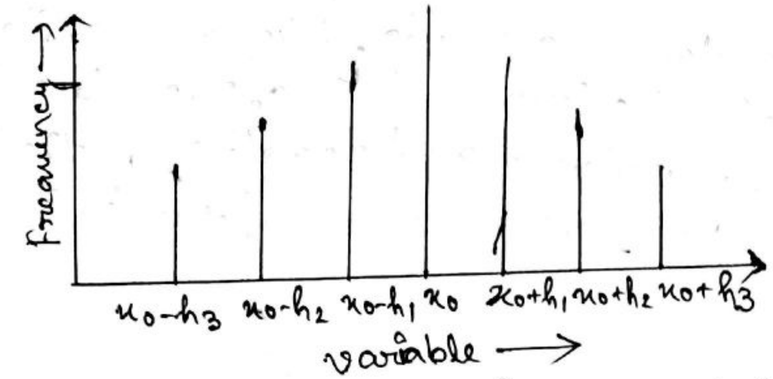
ii) The range of the variable is finite and there is high-order contact at the terminal of the range.

iii) the total frequency should be fairly large.

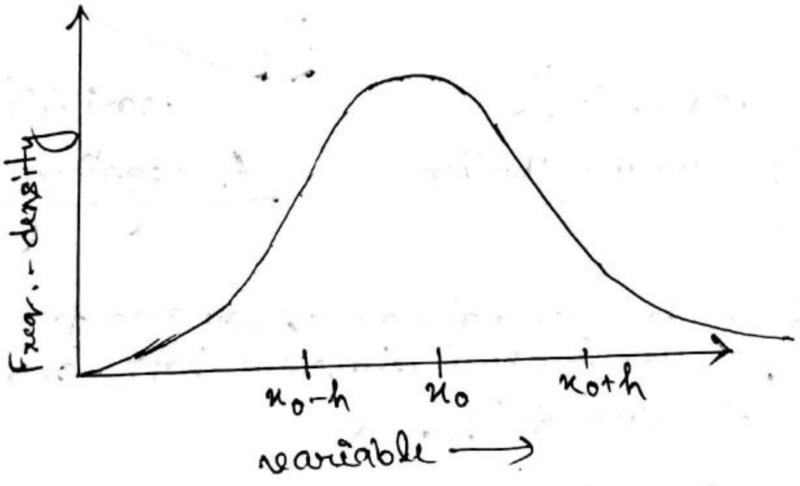
iv) the width of the classes should not be too small in comparison with the range of variation of the data, i.e., the number of classes should not be too large.

⇨ Give the defn of Symmetric distribution →

A frequency distribution of a discrete variable x is called symmetric about x_0 if $f(x_0-h)=f(x_0+h)$, for all h, where $f(x)$ is the freq. of x.



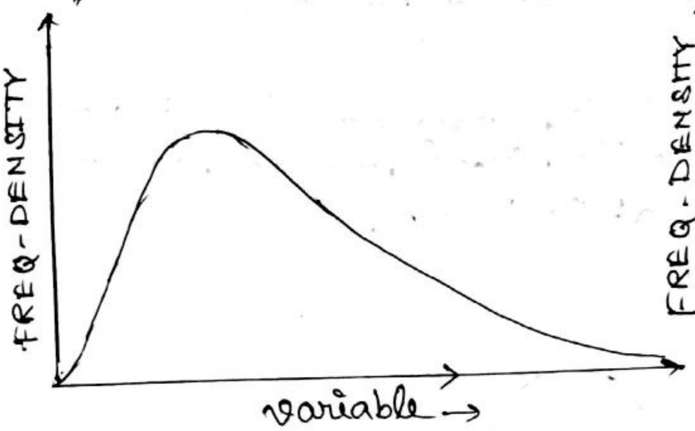
For a continuous variable, a frequency distn. is said to be symmetric about x_0 if $f(x_0-h)=f(x_0+h)$, for all h, where $f(x)$ is the frequency-density at the point x.



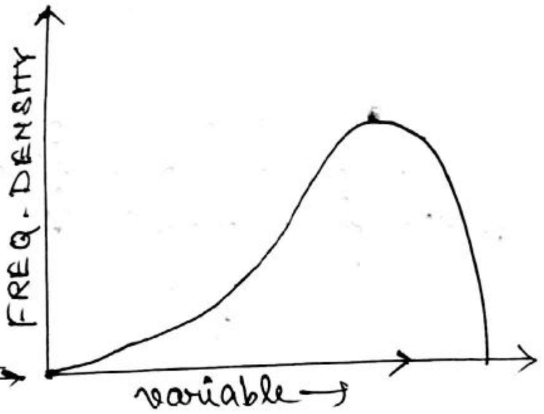
⇨ Asymmetrical or skew distribution →

If a distribution is not found to be symmetrical, then it is termed as skew or asymmetrical. By skewness of a distribution, we mean its degree of deviation / departure from symmetry. The skewness is called positive or negative according as the longer tail of the distribution is towards the higher or the lower values of the variable.

SKEWNESS:→ Skewness means lack in symmetry. By skewness of a freq. distn. we mean the degree of its departure or deviation from symmetry. Skewness indicates whether the freq. curve is inclined more to one side than to the other, i.e. whether the curve has a longer tail on one side. Skewness is said to be positive if the longer tail of the distribution is towards the higher values and negative if the longer tail is towards the lower values of the variable.



A positively skew distribution



A negatively skew distribution

(*) All odd-order central moments are zero for a symmetrical distn., positive for positively skew-distn. and negative for a negatively skew-distn..

Alternative Proof :→

Consider a frequency distribution :

Values	Freq.
$x_0 - h_k$	f_k
$x_0 - h_{k-1}$	f_{k-1}
$x_0 - h_{k-2}$	f_{k-2}
⋮	
$x_0 - h_1$	f_1
x_0	f_0
$x_0 + h_1$	f_1
⋮	
$x_0 + h_{k-1}$	f_{k-1}
$x_0 + h_k$	f_k
Total	n

Note that —

$$m'_{2r-1}(x_0) = \frac{1}{n}\sum_i (x_i - x_0)^{2r-1} f_i$$

$$= \frac{1}{n}\Big\{ (x_0 - h_k - x_0)^{2r-1} f_k + (x_0 - h_{k-1} - x_0)^{2r-1} f_{k-1}$$

$$+ \cdots\cdots + (x_0 - h_1 - x_0)^{2r-1} f_1 + (x_0 - x_0)^{2r-1} f_0$$

$$+ (x_0 + h_1 - x_0)^{2r-1} f_1 + \cdots\cdots + (x_0 + h_{k-1} - x_0)^{2r-1} f_{k-1}$$

$$+ (x_0 + h_k - x_0)^{2r-1} f_k \Big\}$$

$$= \frac{1}{n}\Big\{ (-h_k)^{2r-1} f_k + (-h_{k-1})^{2r-1} f_{k-1} + \cdots + (-h_1)^{2r-1} f_1$$

$$+ 0 + (h_1)^{2r-1} f_1 + \cdots\cdots + h_{k-1}^{2r-1} f_{k-1} + h_k^{2r-1} f_k \Big\}$$

$= 0$, since $2r-1$ is odd and $(-h_i)^{2r-1} f_i = -h_i^{2r-1} f_i$ $\forall\, i$

For $r = 1$,

$$\frac{1}{n}\sum_i (x_i - x_0) f_i = 0$$

$$\Rightarrow \frac{1}{n}\sum_i x_i f_i - x_0 = 0$$

$\Rightarrow x_0 = \bar{x}$ = the A.M of the freq. distn.

Hence,

$$m_{2r-1} = \frac{1}{n}\sum_i (x_i - \bar{x})^{2r-1} f_i$$

$$= \frac{1}{n}\sum_i (x_i - x_0)^{2r-1} f_i$$

$$= 0 , \quad \forall\; r = 1, 2, 3, \ldots\ldots$$

Hence, all odd number central moments of a symmetrical freq. distn. are zero. $\boxed{\text{C.V}}$

⇒ Different measures of Skewness :- 6.2

There are different measures of skewness based on —

i) the moments of the distribution.

ii) the relative positions of the mean, median and mode of the distn.

iii) the relative positions of the quartiles of the distributions.

[C.V]

a) i) Measure based on moments :→ For a symmetric distribution the (2r+1)th central moment $m_{2r+1} = 0$.
So, the least non trivial value of r ; i.e. r = 1. We can use m_3 as a measure of skewness.
But to make our measure of skewness unit free, we use

$$g_1 = \frac{m_3}{s^3},$$

i.e. $\boxed{g_1 = \frac{m_3}{m_2^{3/2}}}$ — as a unit-free measure of skewness.

Therefore, $\begin{cases} g_1 = 0 \Rightarrow \text{the dist}^n \text{ i.e. symmetric.} \\ g_1 > 0 \Rightarrow \text{positively skewed.} \\ g_1 < 0 \Rightarrow \text{negatively skewed.} \end{cases}$

Sometimes, we take → $b_1 = \frac{m_3^2}{m_2^3}$ as a measure of skewness.

when $\begin{cases} b_1 = 0 \Rightarrow \text{the dist}^n \text{ i.e. symmetric} \\ b_1 > 0 \Rightarrow \text{the dist}^n \text{ i.e. asymmetric.} \end{cases}$

[C.V]

ii) Measure of skewness based on the relative position of the Mean, Median, Mode :→ For a moderately skewed distribution it can be verified empirically that —

Mean($\bar{x}$) > Median($\tilde{x}$) > Mode($\breve{x}$) : when the dist^n. is positively skewed.

Mean($\bar{x}$) < Median($\tilde{x}$) < Mode($\breve{x}$) : when the dist^n. is negatively skewed.

Mean($\bar{x}$) = Median($\tilde{x}$) = Mode($\breve{x}$) : for symmetric dist^n.

Considering this characteristics of a freq. dist^n.
Pearson suggested two measures of skewness —

$$S_{k_1} = \frac{\text{Mean} - \text{Mode}}{S_x} \quad ; \quad S_{k_2} = \frac{3(\text{Mean} - \text{Median})}{S_x}$$

↳ Pearson's 1st measure of skewness; ↳ Pearson's 2nd measure of skewness

The second measure was suggested on observing that for a moderately skewed dist^n. we have —

$$\text{Mean} - \text{Mode} \simeq 3(\text{Mean} - \text{Median})$$

[Empirical relation]

[C.U]

Result :→ $|Sk_2| \le 3$

Proof :→ We note that — $|\bar{x}-\tilde{x}|$

$$= \left|\frac{1}{n}\sum_{i=1}^{n} x_i - \tilde{x}\right|$$

$$\le \frac{1}{n}\sum_{i=1}^{n} |x_i - \tilde{x}|$$

$$\le \frac{1}{n}\sum |x_i - \bar{x}|$$

$$\le \sqrt{\frac{1}{n}\sum (x_i-\bar{x})^2} = S_x$$

$$\therefore \frac{|\bar{x}-\tilde{x}|}{S_x} \le 1 \quad \text{, i.e. } |Sk_1| \le 1$$

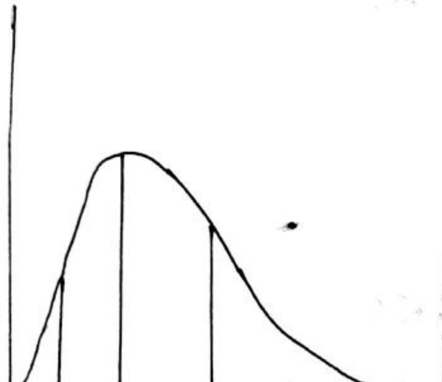
$$\Rightarrow Sk_2 = \frac{3|\bar{x}-\tilde{x}|}{S_x} \le 3$$

$$\Rightarrow |Sk_2| \le 3$$

Hence, the measure of skewness given by the relative position of AM and median must lies between −3 and 3.

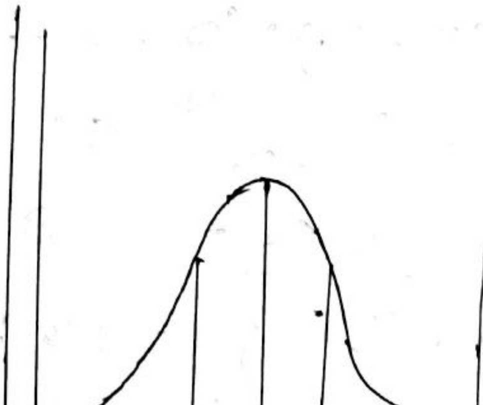
[C.U]

iii) Measure based on relative position of quartiles :→



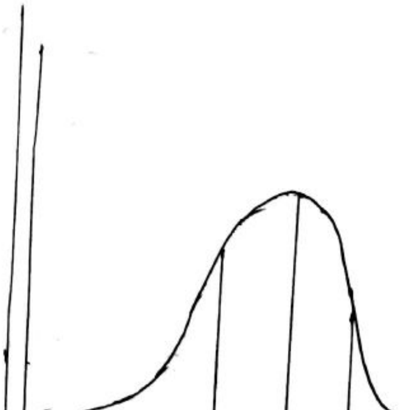
(positively skewed distn.)

$[(Q_3-Q_2)-(Q_2-Q_1) > 0]$



(Symmetric distribution)

$[(Q_3-Q_2)-(Q_2-Q_1) = 0]$



(negatively skewed distn.)

$[(Q_3-Q_2)-(Q_2-Q_1) < 0]$

Considering the relative position of Q_1, Q_2 and Q_3 for symmetric and skewed distribution Bowley suggested a measure of skewness—

$$Sk_3 = \frac{(Q_3-Q_2)-(Q_2-Q_1)}{(Q_3-Q_2)+(Q_2-Q_1)} = \frac{Q_3-2Q_2+Q_1}{Q_3-Q_1}$$

— known as Bowley's measure.

C.V

Result:→ $|S_{K_3}| < 1$

Proof:→ To show that $|S_{K_3}| < 1$, we proceed as follows.

$$2(Q_3 - Q_2) > 0$$

$$\Rightarrow (Q_3 - Q_2) > -(Q_3 - Q_2)$$

$$\Rightarrow (Q_3 - Q_2) - (Q_2 - Q_1) > -(Q_3 - Q_2) - (Q_2 - Q_1)$$

$$\Rightarrow (Q_3 - 2Q_2 + Q_1) > -(Q_3 - Q_1)$$

$$\Rightarrow \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1} > -1$$

or, $S_{K_3} > -1$. ——— ①

Again we proceed as follows —

$$2(Q_2 - Q_1) > 0$$

$$\Rightarrow (Q_2 - Q_1) > -(Q_2 - Q_1)$$

$$\Rightarrow (Q_3 - Q_2) + (Q_2 - Q_1) > (Q_3 - Q_2) - (Q_2 - Q_1)$$

$$\Rightarrow (Q_3 - Q_1) > (Q_3 - 2Q_2 + Q_1)$$

$$\Rightarrow \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1} < 1$$

or, $S_{K_3} < 1$ ——— ②

So from ① and ②, we get — $|S_{K_3}| < 1$

Hence, the measure of skewness given by the relative position of quartiles must lie between -1 and 1.

KURTOSIS :→

6.2

Ques :→(a) What do you mean by kurtosis of a freq. distn ? Distinguish between the leptokurtic, mesokurtic and platykurtic distns. (5)

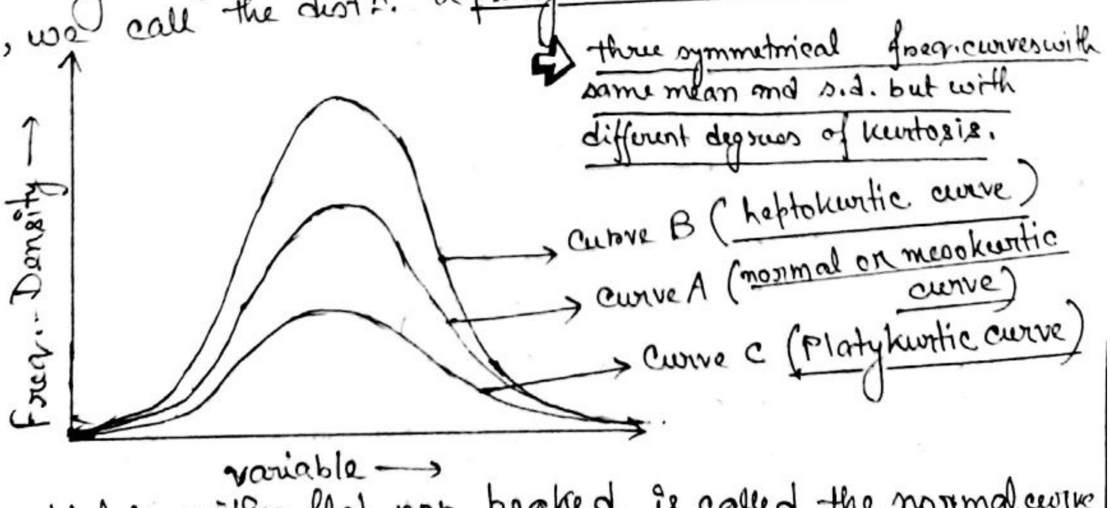
(b) What can you say about the tails of the above three distributions when they are symmetric and have same variance? (3)

(c) Give a measure of kurtosis using moments. When this measure will be appropriate. (4)

Ans :→ C.U

(a) ☒ Kurtosis of a freq. distn. measures the degree of peakedness of a distribution. Two distribution may have the same mean and the same standard deviation and may be equally skew, but one of them may be more peaked than the other. Actually A frequency curve may be symmetrical but it may fail to be equally flat, topped with normal curve. kurtosis enables us to have an idea about the flatness ~~and~~ or peakedness of the freq. curve. So, the relative "flatness or peakedness" of the top of a freq. curve is called the kurtosis.

☒ If the peak of the distribution is moderately high then we call the distn. a Mesokurtic distribution. If the height of the peak of the freq. distn. is more than this moderate height we call the distn. a leptokurtic distribution. On the other hand if the height of the peak of the distn. is less than this moderate height, we call the distn. a platykurtic distribution.



three symmetrical freq. curves with same mean and s.d. but with different degrees of kurtosis.

Freq.- Density →
variable →
Curve B (leptokurtic curve)
curve A (normal or mesokurtic curve)
Curve C (Platykurtic curve)

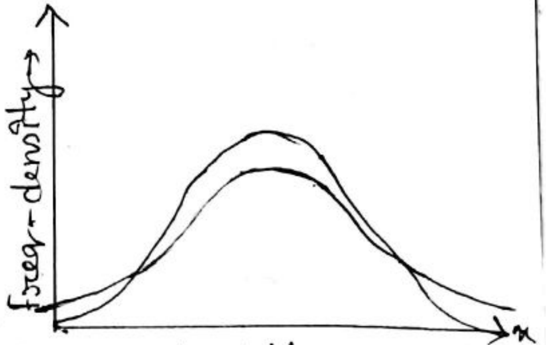
Curve A which is neither flat nor peaked is called the normal curve or mesokurtic curve. Curve C which is flatter than the normal curve is known as. Platykurtic curve. Curve B which is more peaked than the normal curve is called leptokurtic curve.

(b) Leptokurtic distribution has high concentration of values near the central tendency and has high tails, in comparison with a normal distribution with the same standard deviation. Again, Platykurtic distribution has low concentration of values near the average and has low tails, in comparison with a normal distn. with the same standard deviation.

(c) Measures of kurtosis :→

▨ i) Measure of kurtosis using moments :→

A large value of m_4 indicates that the distn. has high concentration of values near the mean and has high tails.



The two freq. distn. given in the figure may have the same m_4 but their flatness of the top are not same. To get a proper measure of kurtosis, we divide m_4 by its order m_2^2 and we get —

$$\boxed{b_2 = \frac{m_4}{m_2^2}}$$ as a measure of kurtosis proposed by Karl Pearson.

Note that b_2 is a unit free measure or a pure number. The value of b_2 for normal curve is 3.

Hence, $g_2 = b_2 - 3 = \frac{m_4}{s^4} - 3 = \frac{m_4}{m_2^2} - 3$. $(s>0)$

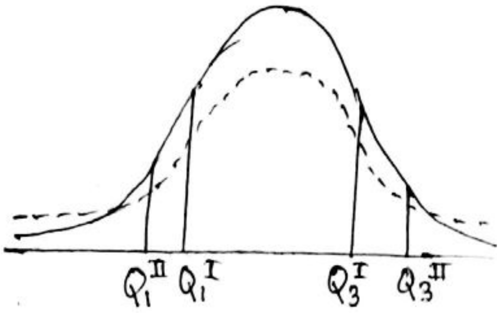
i.e. $\boxed{g_2 = \frac{m_4}{m_2^2} - 3}$ is a measure excess of kurtosis.

For Mesokurtic distn., $g_2 = 0$
For Platykurtic distn., $g_2 < 0$
For Leptokurtic distn., $g_2 > 0$

▨ This measure b_2 (or g_2) will be appropriate as a measure of kurtosis or peakedness only if we confine our attention to the class of the usual bell-shaped (or unimodal) distributions. Otherwise, it may only serve to distinguish a unimodal distribution from a bimodal.

ii) Measure of kurtosis using quartiles :–

$$K_p = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})}$$



Note that, with in the length $Q_3 - Q_1$, we have 50% of the values of a freq. distn..

From the figure, the larger the length $Q_3 - Q_1$, the less the kurtosis of a distribution. Hence, $\frac{Q_3 - Q_1}{2}$ indicates the kurtosis in the negative direction. To get a proper measure of kurtosis, we devide $\frac{Q_3 - Q_1}{2}$ by its order $P_{90} - P_{10}$. therefore, we get –

$$K_p = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})}$$ as a measure of kurtosis.

For Mesokurtic distn., $K_p = 0.263$

For Platykurtic distn., $K_p < 0.263$

For Leptokurtic distn., $K_p > 0.263$

SOME IMPORTANT RESULTS :–

1) $b_2 \geq 1$

Proof :→ Suppose a variable x takes n values $x_1, x_2, \ldots, x_n$ with mean $\bar{x} = \sum_{i=1}^{n} x_i / n$.

In cauchy-schwarz inequality – $\left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right) \geq \left(\sum_{i=1}^{n} a_i b_i\right)^2$,

Let, $a_i = (x_i - \bar{x})^2$, $b_i = 1 \ \forall\ i$, then we get –

$$\left\{\sum_{i=1}^{n}(x_i - \bar{x})^4\right\}\left\{\sum_{i=1}^{n} 1\right\} \geq \left\{\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}^2$$

$$\text{or, } n\sum_{i=1}^{n}(x_i - \bar{x})^4 \geq \left\{\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}^2$$

$$\text{or, } \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4 \geq \left\{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}^2$$

$$\text{or, } m_4 \geq m_2^2$$

$$\text{or, } \frac{m_4}{m_2^2} \geq 1$$

$$\text{or, } b_2 \geq 1$$

'=' holds when the variable takes only two distinct values with same frequency.

2) $b_2 > b_1$

Proof :- Consider C-S inequality —

$$\left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right) \geqslant \left(\sum_{i=1}^{n} a_i b_i\right)^2,$$

we put $a_i = (x_i - \bar{x})^2$ and $b_i = (x_i - \bar{x})$ and we have.

$$\left\{\sum_{i=1}^{n}(x_i-\bar{x})^4\right\}\left\{\sum_{i=1}^{n}(x_i-\bar{x})^2\right\} \geqslant \left\{\sum_{i=1}^{n}(x_i-\bar{x})^3\right\}^2$$

or, $n m_4 \cdot n m_2 \geqslant (n m_3)^2$

or, $m_4 m_2 \geqslant m_3^2$

or, $\dfrac{m_4}{m_2^2} \geqslant \dfrac{m_3^2}{m_2^3}$,

or, $b_2 \geqslant b_1$.

When '=' holds, we must have $x_i - \bar{x} = a$ constant, $\forall$ i, i.e., $x_i = a$ constant; but in that case $s=0$ and, as such b_1 and b_2 are undefined. So $b_2 > b_1$.

$\boxed{\text{C.V}}$

3) $b_2 \geqslant b_1 + 1$

Proof :- Let the variable x assume the values $x_1, x_2, \ldots, x_n$ with mean $\bar{x}$ and standard deviation s.

In cauchy-schwarz inequality —

$$\left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right) \geqslant \left(\sum_{i=1}^{n} a_i b_i\right)^2$$

Let $a_i = \dfrac{x_i - \bar{x}}{s}$, $b_i = \left(\dfrac{x_i - \bar{x}}{s}\right)^2 - 1 = a_i^2 - 1$

Then, by the above inequality —

$$\left[\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^2\right]\left[\sum_{i=1}^{n}\left\{\left(\frac{x_i-\bar{x}}{s}\right)^2 - 1\right\}^2\right] \geqslant \left[\sum_{i=1}^{n}\left\{\left(\frac{x_i-\bar{x}}{s}\right)^3 - \left(\frac{x_i-\bar{x}}{s}\right)\right\}\right]^2$$

$$\Rightarrow \frac{ns^2}{s^2}\left(\frac{nm_4}{s^4} - 2n + n\right) \geqslant \left(\frac{nm_3}{s^3} - 0\right)^2$$

~~or, $n^2\left(\dfrac{m_4}{s^4} - 1\right) \geqslant n^2\left(\dfrac{m_3^2}{s^6}\right)$~~

$$\Rightarrow n^2\left(\frac{m_4}{s^4} - 1\right) \geqslant n^2 \frac{m_3^2}{s^6}$$

$$\Rightarrow \frac{m_4}{m_2^2} - 1 \geq \frac{m_3^2}{m_2^3}.$$

$$\Rightarrow b_2 - 1 \geq b_1$$

$$\Rightarrow \boxed{b_2 \geq b_1 + 1}$$

Equality ('=') holds iff − $a_i \propto b_i$, $\forall i = 1(1)n$

$$\Rightarrow \frac{b_i}{a_i} = k$$

iff $\left(\frac{x_i - \bar{x}}{s}\right)^2 - 1 = k\left(\frac{x_i - \bar{x}}{s}\right)$, $\forall i$

iff $\frac{x_i - \bar{x}}{s} = \pm k$, $\forall i$

iff $x_i = \bar{x} \pm ks$, $\forall i = 1(1)n$.

iff the variable takes only two distinct values [not necessarily with equal frequency].

Examples :-

1. Let $x_1, x_2, \dots, x_n$ be n values of a variable x. Define $u_i = \frac{x_i - \bar{x}}{s}$. Show that $m_3(u)$ and $m_4(u)$ are measures of skewness and kurtosis.

Soln.→ Note that $u_i = \frac{x_i - \bar{x}}{s}$ is a standard value and is independent of order or unit of x. Note that —

$$m_3(u) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3 = \frac{m_3 x}{s^3} = \frac{m_3(x)}{\{m_2(x)\}^{3/2}} = g_1 \text{ and}$$

$$m_4(u) = \frac{m_4(x)}{\{m_2(x)\}^2} = g_2$$ — are the measures of skewness and kurtosis.

2. For a set of values x_i's , $i = 1(1)n$, show that − $\begin{vmatrix} 1 & \Sigma x_i & \Sigma x_i^2 \\ \Sigma x_i & \Sigma x_i^2 & \Sigma x_i^3 \\ \Sigma x_i^2 & \Sigma x_i^3 & \Sigma x_i^4 \end{vmatrix} \geq 0$.

Hence, show that — $g_2 \geq g_1 + 1$.

Soln.→ $\sum_{i=1}^{n}(a_0 + a_1 x_i + a_2 x_i^2)^2 \geq 0$ for all (a_0, a_1, a_2)

$$\Leftrightarrow a_0^2 + a_1^2(\Sigma x_i^2) + a_2^2(\Sigma x_i^4) + 2a_0 a_1(\Sigma x_i) + 2a_0 a_2(\Sigma x_i^2) + 2a_1 a_2(\Sigma x_i^3) \geq 0, \ \forall (a_0, a_1, a_2)$$

Hence, the L.H.S is a quadratic form in (a_0, a_1, a_2) and it is n.n.d.

Hence, $\begin{vmatrix} 1 & \Sigma x_i & \Sigma x_i^2 \\ \Sigma x_i & \Sigma x_i^2 & \Sigma x_i^3 \\ \Sigma x_i^2 & \Sigma x_i^3 & \Sigma x_i^4 \end{vmatrix} \geq 0$. Replacing x_i by $(x_i - \bar{x})$; we get

$$\begin{vmatrix} 1 & 0 & nm_2 \\ 0 & nm_2 & nm_3 \\ nm_2 & nm_3 & nm_4 \end{vmatrix} \geq 0 \Rightarrow g_2 \geq g_1 + 1.$$

Q.Show that $b_2 \geq b_1+1$ by Matrix method.

ANS:-

Considering,

$$\sum_{i=1}^{n} (a_0+a_1x_i+a_2x_i^2)^2 \geq 0 \quad \forall (a_0, a_1, a_2)$$

$$\Rightarrow \sum_{i=1}^{n} [a_0^2 + a_1^2 x_i^2 + a_2^2 x_i^4 + 2a_0a_1x_i + 2a_0a_2x_i^2 + 2a_1a_2x_i^3] \geq 0$$

$$\Rightarrow \sum_{i=1}^{n} a_0^2 + a_1^2 \sum_i x_i^2 + a_2^2 \sum_i x_i^4 + 2a_0a_1\sum_i x_i + 2a_0a_2\sum_i x_i^2 + 2a_1a_2\sum_i x_i^3 \geq 0$$

$$\Rightarrow n a_0^2 + a_1^2\left(\sum_i x_i^2\right) + a_2^2\left(\sum_i x_i^4\right) + 2a_0a_1\left(\sum_i x_i\right) + 2a_0a_2\left(\sum_i x_i^2\right) + 2a_1a_2\left(\sum_i x_i^3\right) \geq 0$$

$$\Rightarrow \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} \begin{pmatrix} n & \sum_i x_i & \sum_i x_i^2 \\ \sum_i x_i & \sum_i x_i^2 & \sum_i x_i^3 \\ \sum_i x_i^2 & \sum_i x_i^3 & \sum_i x_i^4 \end{pmatrix} (a_0 \quad a_1 \quad a_2) \geq 0$$

As this quadratic form is non-negative definite,

Hence, $$\begin{vmatrix} n & \sum_i x_i & \sum x_i^2 \\ \sum x_i & \sum_i x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{vmatrix} \geq 0$$

Now putting $x_i = x_i - \bar{x}$, we get

$$\begin{vmatrix} n & 0 & \sum(x_i-\bar{x})^2 \\ 0 & \sum(x_i-\bar{x})^2 & \sum(x_i-\bar{x})^3 \\ \sum(x_i-\bar{x})^2 & \sum(x_i-\bar{x})^3 & \sum(x_i-\bar{x})^4 \end{vmatrix} \geq 0$$

we know, $m_r = \frac{1}{n}\sum_i (x_i-\bar{x})^r$

$$\Rightarrow \begin{vmatrix} n & 0 & nm_2 \\ 0 & nm_2 & nm_3 \\ nm_2 & nm_3 & nm_4 \end{vmatrix} \geq 0$$

$$\Rightarrow n^3 m_2m_4 - n^3m_3^2 - n^3m_2^3 \geq 0$$

$\left[b_2 = \frac{m_4}{m_2^2}, \quad b_1 = \frac{m_3^2}{m_2^3} \right]$

$$\Rightarrow m_2m_4 - m_3^2 - m_2^3 \geq 0$$

$$\Rightarrow m_2m_4 \geq m_3^2 + m_2^3$$ [dividing both side by m_2^3]

$$\Rightarrow \frac{m_4}{m_2^2} \geq \frac{m_3^2}{m_2^3} + 1$$

$$\Rightarrow b_2 \geq b_1 + 1$$ [ANSWER]

CHAPTER-5

BIVARIATE FREQUENCY DISTRIBUTIONS

Bivariate Data :– Data on two variables recorded simultaneously for a group of individuals are called bivariate data. Examples of bivariate data are heights and weights of the students in a class, the income and expenditure of a number of families, etc.

Firstly, we want to study the nature and extent of association, if any, between the variables. Secondly, if the variables are found to be associated, we express one of them (regarded as the dependent variable) as a mathematical function of the other (considered as independent variable); so that we can predict the value of the dependent variable when the value of the independent variable is known. The first problem is called correlation analysis and the second the regression analysis.

When there are data for a considerably large number of individuals, they are summarised in a two way frequency table. A suitable number of classes are taken for each variable, keeping in mind the same considerations as in the univariate case. Suppose we are given n pairs of values of variable x and y. If there be k classes for x and l classes for y, the freq. table will have $k \times l$ cells. With the help of tally marks we can find the frequencies of different cells. The whole set of class-frequencies define the Bivariate frequency distribution of variables x and y.

Bivariate frequency Distribution :– If X and Y are two variables, then we can observe them together. If X and Y are observed as paired obsn.s, then we have same no. of obsn.s on X and Y. In that case the obsn. may be recorded as follows:

X	x_1	x_2	. . . .	x_n
Y	y_1	y_2	. . . .	y_n

If X and Y are observed individually, then we may have different no. of obsn.s on X and Y. Let there be k obsn.s. on X, say $x_1, x_2, \ldots, x_k$ and l obsns. on Y, say $y_1, y_2, \ldots, y_l$. Also let, f_{ij} be freq. of individual obsns with $X = x_i$, $Y = y_j$. In that case observation may be recorded in a bivariate freq. table as follows :

X \ Y	(y_0-y_1) y_1	(y_1-y_2) y_2	- - - -	$(y_{j-1}-y_j)$ y_j	- - - - - -	$(y_{l-1}-y_l)$ y_l	Total
$(x_0-x_1)\mid x_1$	f_{11}	f_{12}	- - - -	f_{1j}	- - - - - -	f_{1l}	f_{10}
$(x_1-x_2)\mid x_2$	f_{21}	f_{22}	- - - -	f_{2j}	- - - - - -	f_{2l}	f_{20}
⋮	⋮	⋮		⋮		⋮	⋮
$(x_{i-1}-x_i)$ x_i	f_{i1}	f_{i2}	- - - - - -	f_{ij}	- - - - - - - -	f_{il}	f_{i0}
⋮	⋮	⋮		⋮		⋮	⋮
$(x_{k-1}-x_k)$ x_k	f_{k1}	f_{k2}	- - - -	f_{kj}	- - - - - - -	f_{kl}	f_{k0}
TOTAL	f_{01}	f_{02}	- - - - -	f_{0j}	- - - - - - - -	f_{0l}	n

where, $f_{i0}=\sum_{j=1}^{l} f_{ij}$ for $i=1(1)k$.

$=$ Freq. of $X = x_i$, irrespective of the value taken by Y.

Similarly, $f_{0j}=\sum_{i=1}^{k} f_{ij}$ for $j=1(1)l$.

$=$ Freq. of $Y = y_j$, irrespective of the value taken by X.

Here, $f_{i0} = (i=1(1)k)$ and $f_{0j}\,(j=1(1)l)$ are called the marginal frequencies of $X=x_i\,(i=1(1)k)$ and $Y=y_j\,(j=1(1)l)$ and —

$$n=\sum_{i=1}^{k} f_{i0}=\sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij}=\sum_{j=1}^{l}\left(\sum_{i=1}^{k} f_i\right)=\sum_{j=1}^{l} f_{0j}$$

For this bivariate freq. table, the means and the variance of X and Y may be defined as ——

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{k} x_i f_{i0}\,,\quad \bar{Y}=\frac{1}{n}\sum_{j=1}^{l} y_j f_{0j}\,,$$

$$s_X^2=\frac{1}{n}\sum_{i=1}^{k}(x_i-\bar{x})f_{i0}\,,\quad s_Y^2=\frac{1}{n}\sum_{j=1}^{l}(y_j-\bar{y})f_{0j}\,.$$

For such a bivariate freq. table, we can define another additiona
measure,

$$s_{XY}=\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l}(x_i-\bar{x})(y_i-\bar{y})f_{ij}$$ — called co-variance of X and

The co-variance is useful for investigating relationship between X and
Note that if X and Y are observed as paired obs^n. and if they wer
recorded as a non-freq. type data as follows:

X	x_1	x_2	....	x_n
Y	y_1	y_2	....	y_n

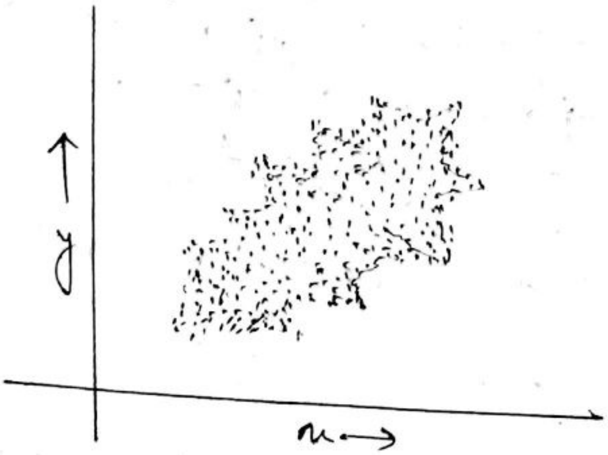
then, $s_{XY}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$

C.U.

SCATTER DIAGRAM (OR DOT DIAGRAM) :→

The simplest mode of diagrammatic representation of bivariate data is the use of a scatter diagram. Suppose we are given n pairs of vealues of variables x and y. Taking two mutually perpendicular straight lines as axes of refference for x and y, each pair of given valeees can be plotted as a point in the graph paper. The figure obtained, when all the n pairs of values have plotted, is called a scatter diagram(or a dot diagram).

From a scatter diagram one can know the nature and the intensity of association, if any, between the variables under study.



Limitation — It is not a suitable method when the number of individuals is very large.

C.U.

▨ **Bivariate Data :—** Most of our discussions so far have been confined to a single variable (univariate data). In statistical work, we often have to deal with problems involving more than one variable. Our interest now lies in studying the relationship between two variables.

Suppose we have data on two variable, say x and y, for each individual in a group e.g., x may be the height and y is the weight of adult male of some athelic group. Our raw data will then consist of a number of pair of values of x and y ; $\{x_i, y_i\}$, $i = 1(1)n$. This type of data is called Bivariate Data.

C.U

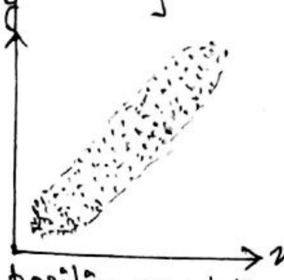
10.1

▨ **CORRELATION :—** If it is found that as one variable increases, the other also increases on the average. there will be said to be positive correlation between two variables. If it is found that as one variable increases, the other variable decreases on the average, we then say that there is a negative correlation between the two variables. There still may be a third situation where as one variable increases, the other remaining constant on the average, this is the case of zero or no correlation.

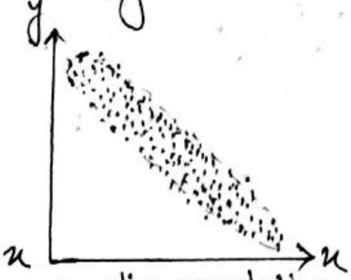
Above consideration are appropriate in case the variables are found to be linearly related, at least in approximate sense.

C.U

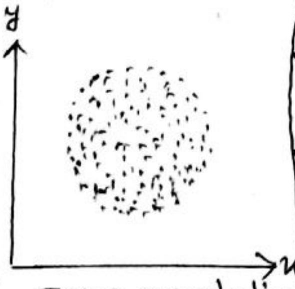
⇨ **Use of Scatter Diagram :—** The scattered diagram serves as a useful technique in the study of the relationship and also for measuring the extent of the linear relationship.



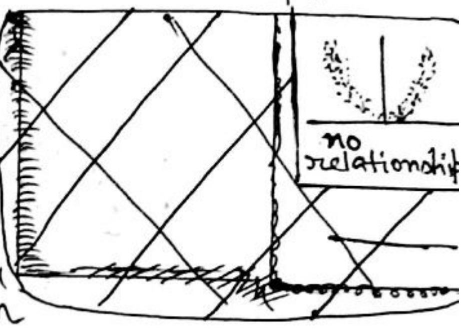
positive correlation



negative correlation



zero correlation



Definition :→ The degree of extent to which the variables are linearly related is called correlation between two variables.

▨ Correlation Coefficient :—

⇒ **Product moment correlation coefficient or Karl Pearson's Coefficient of correlation :—**

It is a measure of linear association between two variables. The correlation coefficient of variables x and y, denoted by r_{xy} (or simply by r when there is no scope of cofusion), is defined as

$$r_{xy} = \frac{\text{cov}(x,y)}{\sqrt{\text{var}(x)}\sqrt{\text{var}(y)}}$$

where cov(x,y) denotes the covariance of x and y.

If we are given n pairs of values (x_i, y_i), $i=1(1)n$, of variables x and y,

$$\text{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}),$$

where $\bar{x}$ and $\bar{y}$ are means of the values of x and y respectively.

$$= \frac{1}{n}\sum_i x_i y_i - \bar{x}\bar{y}$$

$$\text{var}(x) = \frac{1}{n}\sum_i (x_i-\bar{x})^2 = \frac{1}{n}\sum_i x_i^2 - \bar{x}^2$$

and, similarly, $\text{var}(y) = \frac{1}{n}\sum_i y_i^2 - \bar{y}^2$

So, we can write

$$r_{xy} = \frac{\frac{1}{n}\sum_i x_i y_i - \bar{x}\bar{y}}{\sqrt{\frac{1}{n}\sum_i x_i^2 - \bar{x}^2}\sqrt{\frac{1}{n}\sum_i y_i^2 - \bar{y}^2}}$$

$$= \frac{n\sum_i x_i y_i - \left(\sum_i x_i\right)\left(\sum_i y_i\right)}{\sqrt{n\sum_i x_i^2 - \left(\sum_i x_i\right)^2}\sqrt{n\sum_i y_i^2 - \left(\sum_i y_i\right)^2}}$$

[multiplying numerator and denominator by n^2]

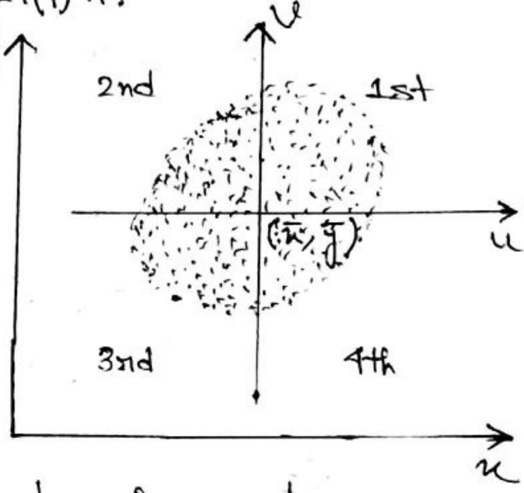
This form is very convenient for computation of correlation coefficient from ungrouped data.

Scanned by CamScanner

Rationale behind taking r as a measure of linear correlation (or correlation) is as follows :

Consider the bivariate data $\{(x_i, y_i) : i=1(1)n\}$

Define, $u_i = x_i - \bar{x}$, $v_i = y_i - \bar{y}$ $\forall$ $i = 1(1)n$.



The origin of the new axis is at $(\bar{x},\bar{y})$, the points on the 1st and 3rd quadrants contribute positive values to $\sum_{i=1}^{n} u_i v_i$ and the points on the 2nd and 4th quadrants contribute negative values to $\sum_{i=1}^{n} u_i v_i$ will indicate a tendency of the points (u_i, v_i) to lie near a line along the 1st and 3rd quadrants, i.e. a strong positive correlation.

Note that, the quantity $\sum_i u_i v_i$ respons to the order of $\sqrt{\sum_i u_i^2} \cdot \sqrt{\sum_i v_i^2}$. To make, the measure suitable for comparitive study, we devide $\sum_{i=1}^{n} u_i v_i$ by $\sqrt{\sum_i u_i^2}\sqrt{\sum_i v_i^2}$ and the resulting measure becomes $\dfrac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_i u_i^2}\sqrt{\sum_i v_i^2}}$ which is independent of the units of measurement of the variables.

Hence, the

$$r_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

$$= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})/n}{s_x s_y} \longrightarrow (*)$$

where s_x and s_y are standard deviations and assumed to be greaters than zero.

$(*)$ is a measure of correlations between x and y and is known as product moment correlation coefficient between x and y.

Covariance :→ Covariance between x and y is defined as

$$\text{cov}(x,y) = \frac{1}{n}\sum_i (x_i-\bar{x})(y_i-\bar{y})$$

$$= \frac{1}{n}\sum_i x_i y_i - \bar{x}\bar{y}.$$

Note that, $\text{cov}(x,x) = \frac{1}{n}\sum_i (x_i-\bar{x})^2 = S_x^2$

$$\text{cov}(y,y) = S_y^2$$

$\text{cov}(x,a) = \frac{1}{n}\sum (x_i-\bar{x})(a-a) = 0$ [∵ a is constant]

$$\text{Cov}(x+y,z) = \frac{1}{n}\sum_i (x_i+y_i-\bar{x}-\bar{y})(z_i-\bar{z})$$

$$= \frac{1}{n}\sum_i (x_i-\bar{x})(z_i-\bar{z}) + \frac{1}{n}\sum_i (y_i-\bar{y})(z_i-\bar{z})$$

10.2 $\qquad = \text{cov}(x,z) + \text{cov}(y,z).$

Definition :→ The product moment correlation coefficient between two variables x and y is defined as

$$r_{xy} = \frac{\text{cov}(x,y)}{\sqrt{\text{var}(x)}\sqrt{\text{var}(y)}}$$, as a measure of linear relationship between the variables x and y.

▨ Properties of Correlation Coefficient :— 10.2

1) The correlation coefficient between x and y is a pure number and is independent of the units of the measurement of x and y.

2)* The correlation co-efficient is independent of the origin and the scale of the variables.

Proof :→ Let, $U_i = \frac{x_i - a}{c}$ and $V_i = \frac{y_i - b}{d}$

where a, b, c, d are arbitrary constants, and $c, d \neq 0$. then ~~for the~~ corresponding to each given pairs of values (x_i, y_i) of x and y, we have a pair of values (U_i, V_i) of U and V, So, we have, $x_i = a + cU_i$ and $y_i = b + dV_i$.

Thus, $\text{var}(x) = \frac{1}{n}\sum_i (x_i-\bar{x})^2 = \frac{c^2}{n}\sum_i (u_i-\bar{u})^2 = c^2\,\text{var}(U)$.

Similarly, $\text{var}(y) = d^2\,\text{var}(V)$.

and, $\text{cov}(x,y) = \frac{1}{n}\sum_i (x_i-\bar{x})(y_i-\bar{y}) = \frac{cd}{n}\sum (u_i-\bar{u})(v_i-\bar{v})$

$$= cd\,\text{cov}(u,v).$$

Hence, $r_{xy} = \dfrac{cov(x,y)}{\sqrt{var(x)}\,\sqrt{var(y)}}$

$$= \frac{cd\ cov(u,v)}{\sqrt{c^2 var(u)}\,\sqrt{d^2 var(v)}}$$

$$= \frac{cd.\ cov(u,v)}{|c|.|d|\sqrt{var(u)}\sqrt{var(v)}}$$

$$= \frac{cd}{|c|.|d|}\, r_{uv}$$

$$= \begin{cases} r_{uv}, & \text{when } c \,\&\, d \text{ are of the same sign.} \\ -r_{uv}, & \text{when } c \,\&\, d \text{ are of the opposite sign.} \end{cases}$$

3) Correlation coefficient r_{xy} of variable x and y is symmetric in x and y, i.e. $r_{xy} = r_{yx}$.

4) $|r_{xy}| \le 1$, or, $-1 \le r\,(\text{or}, r_{xy}) \le 1$ 10.2

Proof:→ ▨ Cauchy-schwartz inequality :

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \le \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right)$$

Take $a_i = x_i - \bar{x}$ and $b_i = y_i - \bar{y}$;

then, $\left\{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right\}^2 \le \left\{\sum_{i=1}^{n}(x_i-\bar{x})^2\right\}\left\{\sum_{i=1}^{n}(y_i-\bar{y})^2\right\}$

$$\Rightarrow \left\{\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2}\sqrt{\Sigma(y_i-\bar{y})^2}}\right\}^2 \le 1$$

$$\Rightarrow r_{xy}^2 \le 1$$

$$\Rightarrow -1 \le r_{xy} \le 1.$$

iff '=' holds in c-s inequality,

iff $(y_i-\bar{y}) = k(x_i-\bar{x})$ $\forall\, i$

iff the variables x and y are linearly related at least for the given data. and

if $r_{xy} = +1$ (or -1), then two variables are exactly linearly related with positive (or negative) slope.

i.e. If $r_{xy} = +1$ then $(y_i-\bar{y}) = k(x_i-\bar{x})$ with $k>0$

$r_{xy} = -1$ then $(y_i-\bar{y}) = k(x_i-\bar{x})$ with $k<0$.

Ex. → Consider the data :

x	-3	-2	-1	0	1	2	3
y	9	4	1	0	1	4	9

Calculate r_{xy}. Hence comment on the relationship between x and y.

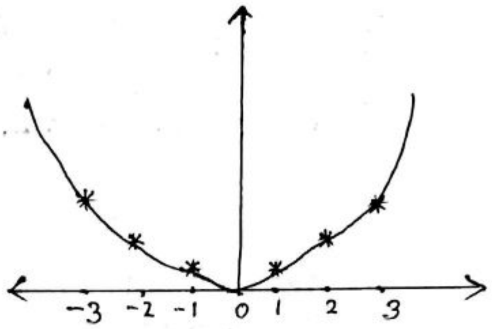
Soln. → Here, $\bar{x}=0$, $\sum x_i y_i = 0$

Hence, $cov(x,y) = \frac{1}{n}\sum x_i y_i - \bar{x}\bar{y} = 0$

$\Rightarrow r_{xy} = 0$

Note that $r_{xy}=0 \Rightarrow$ there is no linear relationship but there may be any non-linear relationship.

Clearly, there is a quadratic relationship such as $y=x^2$. Hence, $r_{xy}=0$, the variables are not independent. There is a perfect dependence of y on x i.e. $y=x^2$. there is no linear relationship. That is why the correlation coefficient is zero.



scatter diagram

REGRESSION : $\underset{=}{10.1}$ Consider the bivariate data (x_i, y_i), $i=1(1)n$ on two variables x and y. Suppose n pairs of values are arranged in arrays of y values corresponding to fixed values of x.

x - values	y - values	means
x_1	y_{11} y_{12} $\cdots\cdots$ y_{1n_1}	$\bar{y}_{x_1}$
x_2	y_{21} y_{22} $\cdots\cdots$ y_{2n_2}	$\bar{y}_{x_2}$
$\vdots$	$\vdots$ $\vdots$ $\vdots$	$\vdots$
x_K	y_{K1} y_{K2} $\cdots\cdots$ y_{Kn_K}	$\bar{y}_{x_K}$

Define $\bar{y}_{xi} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$, $i=1(1)K$, as the mean of the ith array or as the conditional mean of y when $x=x_i$ is given. If we plot the array mean $\bar{y}_n$ for different values of x in a graph-paper, the curve obtained by the points $(x, \bar{y}_n)$ is called the regression curve and its equation is called the regression equations of y on x. Clearly, $\bar{y}_n$ is a function of x and the equation $y=\bar{y}_n$ is called the regression equation of y on x.

● By regression of a variable y on another variable x we mean the dependence of y on x, on the average. In bivariate analysis, one of the major problem is prediction of the value of the dependent variable y when the value of the independent variable x is known. $y = f(x)$ equation is called the regression equation where y is a mathematical function of x.

Linear & non-linear Regression :→ If the regression of y on x is a linear function of x, i.e. $y = a+bx$, then we say that the regression of y on x is linear, otherwise it is non-linear.

⇨ Regression Analysis :→ Consider the problem of prediction of the value of one variable for the given value of another variable. To solve this problem, it is necessary to develop a relation between the variables x and y, at least in approximate sense.

The regression analysis is a method of finding an average relationship between the variables under study. If we wish to predict the value of y for a given values of x, then y is known as dependent or predicted variable and x, the independent or predictor variable.

Let $y = f(x)$ be a prediction of y value. Consider the model: $y = f(x) + e$, where e is the error in prediction.

Result :→ For an array data $\{(x_i, y_{ij}) : i = 1(1)k,\ j = 1(1)n_i\}$, show that $\bar{y}_x$, the array mean is the best predictor in the sense of having minimum error sum of squires (SSE).

Proof :→ Let $y = f(x)$ be a predictor of y.
Our model is $y = f(x) + e$.
Here $y_{ij} = f(x_i) + e_{ij}$ $\forall$ $j = 1(1)n_i$, $i = 1(1)k$.

Values of x	Values of y	Array means
x_1	$y_{11}\ y_{12}\ \cdots\cdots\ y_{1n_1}$	$\bar{y}_{x_1}$
x_2	$y_{12}\ y_{22}\ \cdots\cdots\ y_{2n_2}$	$\bar{y}_{x_2}$
$\vdots$	$\vdots\ \ \vdots\ \ \vdots$	$\vdots$
x_k	$y_{k1}\ y_{k2}\ \cdots\ y_{kn_k}$	$\bar{y}_{x_k}$

To find the best prediction, we have to minimize error sum of squares $\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\{y_{ij}-f(x_i)\}^2$

Now, $\sum_{i=1}^{k}\sum_{j=1}^{n_i}\{y_{ij}-f(x_i)\}^2$

$$= \sum_i \sum_j \{y_{ij} - \bar{y}_{ni} + \bar{y}_{ni} - f(x_i)\}^2$$

$$= \sum_i\sum_j (y_{ij}-\bar{y}_{ni})^2 + \sum_i\sum_j(\bar{y}_{ni}-f(x_i))^2 + 2\sum_i\sum_j(y_{ij}-\bar{y}_{ni})(\bar{y}_{ni}-f(x_i))$$

$$= \sum_i\sum_j (y_{ij}-\bar{y}_{ni})^2 + \sum_i n_i(\bar{y}_{ni}-f(x_i))^2$$

[∵ product term $= 2\sum_j(y_{ij}-\bar{y}_{ni})\sum_i(\bar{y}_{ni}-f(x_i))$

$= 2\sum_i(\bar{y}_{ni}-f(x_i)) \times 0 = 0$

since $\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{ni}) = 0$]

Hence, $\sum_i\sum_j e_{ij}^2 = \sum_i\sum_j (y_{ij}-\bar{y}_{ni})^2 + \sum_i n_i\{\bar{y}_{ni}-f(x_i)\}^2$

$\geq \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{ni})^2$ for any $f(\cdot)$.

and $\sum_i\sum_j e_{ij}^2$ attains its minimum, when

$\sum_{i=1}^{k} n_i\{\bar{y}_{ni}-f(x_i)\}^2 = 0$

$\Leftrightarrow f(x_i) = \bar{y}_{ni} \ \forall\ i$

Hence, $\bar{y}_{ni}$ is the best prediction of y in the sense of having minimum SSE.

▨ Therefore, by virtue of the above result, we shoulld take our model as

$y = \bar{y}_n + e$ — is Regression analysis.

Least Square Regression (Approximate Regression) :→

In regression analysis, we take our model as $y = \bar{y}_n + e$. the regression curve of y on x may be a curve of the complecated form then we wish to approximate the regression curve by a simple curve. Ex.→ a line, a parabola, etc. Then we write $y = g(x) + \{\bar{y}_n - g(x) + e\}$, Here $y = g(x)$ is a simple approximation of $y = \bar{y}_n$.

Here e is error, a random quantity with mean 0 and constant variance. But $z = \{\bar{y}_n - g(x) + e\}$ is known as residual quantity, the part in y after considering $y = g(x)$ as a predicting formula. The function $y = g(x)$ is determined by minimizing the SSE [or, Residual sum of squares (RSS)] $= \sum_i \{y_i - g(x_i)\}^2$

The equation $y = g(x)$ obtained by the method of least square is known as least square regression equation of y on x.

▨ Regression Equation and Least square linear Regression :→

If for a given set of paired observation $\{(x_i, y_i) : i = 1(1)n\}$ the correlation coefficient $|r_{xy}|$ is quite high (i.e. close to 1), then it indicates that there is a near linear relationship between x and y. To estimate that relationship we fit a line on the observed set of data by the principle of least square as follows :

If Y_i is the regression estimate of y_i, then under the assumption of linear relationship between x and y, our predicting formula is : $Y_i = a + bx_i$.

Hence, our model is : $y_i = Y_i + e_i$, where Y_i is the predicted value of y_i and e_i is the residual or error in the prediction when $x = x_i$

Here the intercept 'a' and the 'slope' b of the regression equation is estimated by the method of least squares which consists of minimizing the error sum of squares (SSE).

$$\sum_{i=1}^{n} e_i^2 = S^2 = \sum_{i=1}^{n}(y_i - Y_i)^2 = \sum_{i=1}^{n}(y_i - a - bx_i)^2$$ is minimum.

w.r.t. a and b. The constants are determined by the equations, called the normal equations.

Now, $\frac{\partial S^2}{\partial a} = 0 \Rightarrow \sum_{i=1}^{n}(y_i - a - bx_i) = 0$

$\Rightarrow \bar{y} = a + b\bar{x}$ — ①

$\frac{\partial S^2}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n}(y_i - a - bx_i)x_i = 0$

$\Rightarrow \sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2$

— these are called normal equations. —— ②

Now, $\{② - n\bar{x}\,①\}$, we get —

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = b\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

$$\Rightarrow \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = b\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\Rightarrow b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\text{cov}(x,y)}{\text{var}(x)} = \frac{r_{xy}\cdot s_x \cdot s_y}{s_x^2}$$

$$\therefore \boxed{b = r_{xy}\cdot\frac{s_y}{s_x}}$$ —— ①

Then from ①, $a = \bar{y} - b\bar{x}$

$$\Rightarrow \boxed{a = \bar{y} - r_{xy}\cdot\frac{s_y}{s_x}\cdot\bar{x}}$$

Hence, the least squares linear regression of y on x is

$$y = \bar{y} + r_{xy}\cdot\frac{s_y}{s_x}(x - \bar{x}) \quad \text{——} (*)$$

The line (*) is known as the regression line of y on x. and in ①, $b_{yx} = b = r_{xy}\cdot\frac{s_y}{s_x}$ is called the regression coefficient of y on x.

when, $x = 0$, $Y = a$

$x = 1$, $Y = a + b$

$x = 2$, $Y = a + 2b$

Hence b_{yx} is the amount by which the predicted value Y increases for a unit increment in x.

Similarly, the least square linear regression of x on y is $x = \bar{x} + r_{xy}\frac{s_x}{s_y}(y - \bar{y})$ —— (**)

NOTE :- The two lines, however, coincide when $r = \pm 1$, i.e. when the relation between the two variables is exactly linear.

⇨ Properties or Results : →

1) Mean of the errors in prediction is 0. the error is uncorrelated with the predictor variable and hence with the predicted values. Mathematically $\bar{e} = 0$, $\text{cov}(e, x) = 0$ and $\text{cov}(e, y) = 0$.

Proof :→ For the normal eauation :

$$\sum_{i=1}^{n} e_i = 0 \quad \text{and} \quad \sum_{i=1}^{n} e_i x_i = 0$$

$$\Rightarrow \bar{e} = 0 \quad \text{and} \quad \text{cov}(e, x) = \frac{1}{n}\sum e_i x_i - \bar{e}\bar{x}$$
$$= 0$$

Now,
$$\text{cov}(e, y) = \text{cov}(e, a + bx)$$
$$= \text{cov}(e, a) + b\,\text{cov}(e, x)$$
$$= 0 + b.0$$
$$= 0 .$$

Remark :⇒ i) we have, $y_i = Y_i + e_i$,

$$\Rightarrow \bar{y} = \bar{Y} + \bar{e} \quad [\text{when } \bar{e} = 0]$$
$$\Rightarrow \bar{y} = \bar{Y} .$$

ii) Note that, $\text{cov}(x, a) = \frac{1}{n}\sum (x_i - \bar{x})(a - a) = 0$

and $\text{cov}(x, by) = b\,\text{cov}(x, y)$.

$$\therefore \text{Cov}(ax + by, cu + dv) = ac.\text{cov}(x, u) + ad\,\text{cov}(x, v) + bc.\text{cov}(y, u) + bd.\text{cov}(y, v)$$

$$\therefore \text{var}(x + y) = \text{cov}(x + y, x + y)$$
$$= \text{cov}(x, x) + \text{cov}(x, y) + \text{cov}(y, x) + \text{cov}(y, y)$$

$$\therefore \text{var}(x + y) = \text{var}(x) + \text{var}(y) + 2\text{cov}(x, y) \quad \left[\because \text{var}(x) = \text{cov}(x, x)\right]$$

and,
$$\text{var}(x - y) = \text{var}(x) + \text{var}(y) - 2\text{cov}(x, y) .$$

2) Let $u = \frac{x-A}{c}$ and $v = \frac{y-B}{d}$, where $c>0$ and $d>0$. Then the regression coefficient of y on x, denoted by b_{yx} for the sake of definiteness, is

$$b_{yx} = \frac{d}{c} \times b_{uv}$$

Proof:→ $x = A + cu$ and $y = B + dv$

so, $\bar{x} = A + c\bar{u}$; and $\bar{y} = B + d\bar{v}$

∴ $x - \bar{x} = c(u - \bar{u})$ and similarly $y - \bar{y} = d(v - \bar{v})$.

Hence,

$$\begin{aligned} \mathrm{Cov}(x,y) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\ &= \frac{1}{n}\sum_{i=1}^{n} c(u_i - \bar{u})d(v_i - \bar{v}) \\ &= c d \cdot \frac{1}{n}\sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v}) \\ &= c \cdot d \cdot \mathrm{cov}(u,v) \end{aligned}$$

$$\begin{aligned} \mathrm{var}(x) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} c^2(u_i - \bar{u})^2 \\ &= c^2 \cdot \frac{1}{n}\sum_{i=1}^{n}(u_i - \bar{u})^2 \\ &= c^2 \cdot \mathrm{var}(u) \end{aligned}$$

So, $b_{xy} = \frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)} = \frac{d}{c} \cdot b_{uv}$.

3) $\bar{y} = \bar{Y}$, where $\bar{y} = a + b\bar{x}$

Proof:→ $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i)$ $[\because Y_i = a + bx_i]$

$= \frac{1}{n}\sum_{i=1}^{n}(\bar{y} - b\bar{x} + bx_i)$ $[\because \bar{y} = a + b\bar{x}]$

$= \frac{1}{n}\sum_{i=1}^{n}[\bar{y} + b(x_i - \bar{x})]$

$= \bar{y} + \frac{b}{n}\sum_{i=1}^{n}(x_i - \bar{x})$

$[\because \sum_i (x_i - \bar{x}) = 0]$

$= \bar{y}$

So that the mean of the predicted values of y is equal to the mean of the corresponding observed values.

From this, it follows that $\bar{e} = \frac{1}{n}\sum e_i = \frac{1}{n}\sum (y_i - Y_i) = \bar{y} - \bar{Y} = 0$.

Similarly results hold for variable x.

4) $\underset{\text{S.T.}}{}$ $\mathrm{Var}(y) = \mathrm{var}(Y) + \mathrm{var}(e)$ and $r_{yY} = \sqrt{\dfrac{V(Y)}{V(y)}}$.

Proof :→ As $y_i = Y_i + e_i$, $i = 1(1)n$, so, $\bar{y} = \bar{Y} + \bar{e}$.

$$\mathrm{Var}(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(Y_i + e_i - \bar{Y} - \bar{e})^2$$

$$= \frac{1}{n}\sum_{i}(Y_i - \bar{Y} + e_i - \bar{e})^2$$

$$= \frac{1}{n}\sum_{i}(Y_i - \bar{Y})^2 + \frac{1}{n}\sum_{i}(e_i - \bar{e})^2 + 2.\frac{1}{n}\sum_{i}(Y_i - \bar{Y})(e_i - \bar{e})$$

$$= \mathrm{var}(Y) + \mathrm{var}(e) + 2.\mathrm{cov}(Y, e)$$

$$= \mathrm{var}(Y) + \mathrm{var}(e).$$ [by property ①, $\mathrm{cov}(Y, e) = 0$]

$$r_{yY} = \frac{\mathrm{cov}(y, Y)}{\sqrt{\mathrm{var}(y)}\sqrt{\mathrm{var}(Y)}}$$

$$= \frac{\mathrm{cov}(Y+e, Y)}{\sqrt{\mathrm{var}(y)} \cdot \sqrt{\mathrm{var}(Y)}}$$

$$= \frac{\mathrm{cov}(Y, Y) + \mathrm{cov}(e, Y)}{\sqrt{\mathrm{var}(y)}\sqrt{\mathrm{var}(Y)}}$$ [$\because \mathrm{cov}(Y, Y) = \mathrm{var}(Y)$; $\mathrm{cov}(e, Y) = 0$]

$$= \frac{\mathrm{var}(Y)}{\sqrt{\mathrm{var}(y)}\sqrt{\mathrm{var}(Y)}}$$

$$= \sqrt{\frac{\mathrm{Var}(Y)}{\mathrm{var}(y)}} . = |r|.$$

which means that the correlation between y and its 'predicted' value Y must be non-negative and must be numerically the same as the correlation between y and x.

5) $|r| = \frac{s_Y}{s_y}$, $\boxed{\text{or}}$ $r^2 = \frac{Var(Y)}{Var(y)}$.

Proof :→ The least square linear regression of y on x is

$$Y_i - \bar{Y} = r\frac{s_y}{s_x}(x_i - \bar{x})$$

Now,
$$V(Y) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$
$$= \frac{1}{n}\left\{ r^2 . \frac{s_y^2}{s_x^2} \times \sum_i (x_i - \bar{x})^2 \right\}$$
$$= r^2 . \frac{s_y^2}{s_x^2} . \left\{ \frac{1}{n}\sum_i (x_i - \bar{x})^2 \right\}$$
$$= r^2 . \frac{s_y^2}{\cancel{s_x^2}} \times \cancel{s_x^2}$$
$$= r^2 . s_y^2$$
$$= r^2 . var(y)$$
$$\therefore \boxed{r^2 = \frac{var(Y)}{var(y)}} \text{ or, } \boxed{r^2 = \frac{s_Y^2}{s_y^2}}.$$

the quantity r^2 is called the coefficient of determination and it may be used as a measure of usefulness of the linear regression equations as prediction formulae.

$\boxed{\text{OR}}$

We know,
$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \text{ and}$$
$$s_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$
$$= \frac{1}{n}\sum_{i=1}^{n}(a + bx_i - \bar{y})^2 \quad [\text{by } \bar{y} = \bar{Y}]$$
$$= \frac{1}{n}\sum_{i=1}^{n}(\cancel{a} + bx_i - \cancel{a} - b\bar{x})^2 \quad [\text{putting } \bar{y} = a + b\bar{x}]$$
$$= \frac{b^2}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$
$$= b^2 . s_x^2$$
$$= r_{xy}^2 . \frac{s_y^2}{s_x^2} . s_x^2$$
$$= r_{xy}^2 . s_y^2$$
$$\therefore r_{xy}^2 = \frac{s_Y^2}{s_y^2} \quad \text{i.e., } \boxed{|r_{xy}| = \frac{s_Y}{s_y}}.$$

i.e. $\boxed{r^2 = \frac{Var(Y)}{var(y)}}$

6) The residual variance, var(e), is given by —

$$var(e) = var(y)\,[1 - r^2]$$

$$\Rightarrow s_e^2 = s_y^2(1-r^2)$$ and prove that —

$$-1 \le r \le 1.$$

Proof:→ If $y_i = Y_i + e_i$

Then, $\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - Y_i) = \bar{y} - \bar{Y} = 0$ [by $\bar{Y} = \bar{y}$]

So, $var(e) = \frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})^2$

$$= \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(y_i - Y_i)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}[y_i - \bar{y} + \bar{y} - Y_i]^2$$

$[\because Y_i = a + bx_i,\ \bar{y} = a + b\bar{x},\ \therefore Y_i - \bar{y} = b(x_i - \bar{x})]$

$$= \frac{1}{n}\sum_{i=1}^{n}[y_i - \bar{y} - b(x_i - \bar{x})]^2$$

$$= \frac{1}{n}\left[\sum_{i=1}^{n}(y_i-\bar{y})^2 + b^2\sum_{i=1}^{n}(x_i-\bar{x})^2 - 2b\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})\right]$$

$$= s_y^2 + b^2 s_x^2 - 2b.s_{xy}$$

$$= s_y^2 + b^2 s_x^2 - 2b^2 s_x^2$$

$$= s_y^2 - b^2.s_x^2$$

$$= s_y^2 - r^2.\frac{s_y^2}{\cancel{s_x^2}}.\cancel{s_x^2}$$

$$\therefore \boxed{s_e^2 = s_y^2(1-r^2)}$$

So, $var(e) = s_e^2 = s_y^2(1-r^2)$

Since, $s_e^2 \ge 0 \Rightarrow (1-r^2) \ge 0$

$\therefore r^2 \le 1$ or $|r| \le 1$.

$$\therefore \boxed{-1 \le r \le 1}.$$

Remark :→

i) In our present model $y = Y + e$ where Y is the predicted or explained part and e is the corresponding residual or unexplained part of y, due to the use of least square linear regression of y on x. Hence, the standard deviation of e, which is called the standard errors of estimate of y from its linear regression on $x = s_{y.x} = s_e = s_y\sqrt{1-r^2}$.

$\therefore s_e^2 = Var(e) = var(y)(1-r^2) = s_y^2(1-r^2)$ is called the the residual or unexplained variability.

Smaller the residual variability (s_e^2), better the prediction.

ii) $s_e^2 \geq 0 \Rightarrow s_y^2(1-r^2) \geq 0 \Rightarrow r^2 \leq 1$

$\Rightarrow -1 \leq r \leq 1$.

- If $r^2 = 1$, i.e. $r = \pm 1$. $\Rightarrow s_e^2 = 0$

$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} e_i^2 = 0$ as $\bar{e} = 0$,

$\Rightarrow e_i = 0 \quad \forall \ i = 1(1)n$ [⇒ there is no error in prediction]

$\Rightarrow y_i - Y_i = 0$

$\Rightarrow y_i = Y_i$

⇒ the given values are linearly related.

So that, all points in the scatter diagram lie on the regression line. Here the linear regression equation will be the ideal predicting formula for y when x is given.

- If $r = 0$, then $s_e^2 = s_y^2 = s_Y^2 + s_e^2$

$\Rightarrow s_Y^2 = 0$.

$\Rightarrow \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = 0$

$\Rightarrow Y_i - \bar{Y} = 0 \ \forall i$

$\Rightarrow Y_i = \bar{Y} = \bar{y} \ \forall i$

⇒ the least square linear regression has no use in predicting the value of y.

$Y = \bar{y}$ means as the linear regression equation is concerned, the value of x throws no light whatever on the value of y.

7) For the variables x and y such that $s_x > 0$ and $s_y > 0$. Show that — $r = \{var(x+y) - var(x-y)\} / 4s_x . s_y$

10.7

Proof :→ $$Var(x+y) = \frac{1}{n}\sum_{i=1}^{n}(x_i + y_i - \bar{x} - \bar{y})^2$$
$$= \frac{1}{n}\sum_{i=1}^{n}\{(x_i - \bar{x}) + (y_i - \bar{y})\}^2$$
$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 + \frac{2}{n}\sum_{i}(x_i - \bar{x})(y_i - \bar{y})$$
$$= var(x) + var(y) + 2cov(x, y)$$

Similarly, $var(x-y) = var(x) + var(y) - 2cov(x, y)$

So, $var(x+y) - var(x-y) = 4cov(x, y)$

$$\Rightarrow \frac{var(x+y) - var(x-y)}{4 \quad s_x s_y} = \frac{\cancel{4}cov(x, y)}{\cancel{4}s_x s_y}$$ [dividing both sides by $4s_x s_y$]

$$\Rightarrow r = \frac{var(x+y) - var(x-y)}{4s_x s_y}$$

8) The correlation coefficient is the geometric mean of the two regression coefficients, the sign of r_{xy} (or r) being common sign of the regression coefficients.

Proof :→ b_{yx} = Regression coefficient of y on x = $r_{xy} \cdot \frac{s_y}{s_x}$

b_{xy} = Regression coefficient of x on y = $r_{xy} \cdot \frac{s_x}{s_y}$

$\therefore b_{yx} \cdot b_{xy} = r^2_{xy}$

i.e. $r_{xy} = \pm\sqrt{b_{xy} \cdot b_{yx}}$

i.e. $|r| = \sqrt{b_{yx} \cdot b_{xy}}$

9) (C.U.)

The angle between the two regression lines is

$$\tan^{-1}\left\{\frac{1-r^2}{|r|}\cdot\frac{s_x . s_y}{s_x^2+s_y^2}\right\} \qquad \underline{10.8.}$$

Interprete the cases : $r=0$, $r=\pm 1$.

Proof:→ the two regression lines are:

$$y=\bar{y}+r\frac{s_y}{s_x}(x-\bar{x}) \quad , \text{ let } r.\frac{s_y}{s_x}=m_1 \text{ (say)}.$$

$$\text{and } x=\bar{x}+r\frac{s_x}{s_y}(y-\bar{y}) \quad \text{and } \frac{s_y}{r s_x}=m_2 \text{ (say)}.$$

If θ be the acute angle between the two lines,

$$\theta=\tan^{-1}\left|\frac{m_1-m_2}{1+m_1m_2}\right|$$

$$=\tan^{-1}\left|\frac{1-r^2}{r}\cdot\frac{s_x s_y}{s_x^2+s_y^2}\right|$$

The other angle between the two lines is $\pi-\theta$.

* If $r=0$, $\theta=\frac{\pi}{2}$ and the two lines at right angles,

C.U. ∴ The two regression lines are perpendicular to each other.

If $r=\pm 1$, $\theta=0$ and the two regression lines coincide.

Examples:

10) Since AM of a number of positive quantities is greater than or equal to their geometric mean,

$$\frac{|b_{yx}|+|b_{xy}|}{2} \geqslant \sqrt{|b_{yx}||b_{xy}|}=\sqrt{b_{yx}.b_{xy}}, \text{ (since } b_{yx} \text{ \& } b_{xy} \text{ are of the same sign)}$$

$$=\sqrt{r^2}=|r|.$$

Thus, numerical value of r cannot exceed the arithmatic mean of the numerical values of the regression co-efficients.

Again, since b_{yx} and b_{xy} are of the same sign, the above inequalities gives $\left|\frac{b_{yx}+b_{xy}}{2}\right|=\frac{|b_{yx}|+|b_{xy}|}{2}\geqslant |r|$.

Examples :-

1) If $4u = 2x+7$ and $6v = 2y-15$, and regression coefficient of y on x is 3, then find the regression coefficient of v on u.

Ans. → $u = \frac{1}{2}x + 7/4$ and $v = \frac{1}{3}y - 5/2$,

$\therefore \bar{u} = \frac{1}{2}\bar{x} + \frac{7}{4}$ and $\bar{v} = \frac{1}{3}\bar{y} - \frac{5}{2}$.

so that, $u - \bar{u} = \frac{1}{2}(x - \bar{x})$ and $v - \bar{v} = \frac{1}{3}(y - \bar{y})$.

$$\therefore var(u) = \frac{1}{n}\sum_{i=1}^{n}(u-\bar{u})^2$$
$$= \frac{1}{4}\cdot\frac{1}{n}\sum_{i=1}^{n}(x-\bar{x})^2$$
$$= \frac{1}{4} var(x)$$

and
$$cov(u,v) = \frac{1}{n}\sum_{i=1}^{n}(u-\bar{u})(v-\bar{v})$$
$$= \frac{1}{6}\cdot\frac{1}{n}\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})$$
$$= \frac{1}{6} cov(x,y)$$

OR

$u = \frac{2x+7}{4}$, $c = 4$

$v = \frac{2y-15}{6}$, $d = 6$

$b_{yx} = 3$

$b_{uv} = \frac{b_{yx}}{d} \times c$

$= \frac{3}{6} \times 4$

$= 2$.

Hence, $b_{vu} = \frac{cov(u,v)}{var(u)} = \frac{2}{3} b_{yx} = \frac{2}{3} \times 3 = 2$ $[\because b_{yx} = 3]$

2) Out of two lines of regression given by $x + 2y - 5 = 0$ and $2x + 3y - 8 = 0$ which one is the regression line of x on y? Use the equations to find the means of x and y. If the variance of x is 12, calculate the variance of y.

Ans :→ The regression lines are $x + 2y - 5 = 0$ ——— (i)

and $2x + 3y - 8 = 0$ ——— (ii)

Let us assume that eqn. (i) is regression line of x on y and eqn. (ii) is that of y on x.

From (i) and (ii) → $x = -2y + 5$ and $y = -\frac{2}{3}x + \frac{8}{3}$, respectively.

we have, in usual notations, $b_{xy} = -2$ and $b_{yx} = -\frac{2}{3}$.

So, $r^2_{xy} = b_{xy} \cdot b_{yx} = \frac{4}{3} > 1$, so our assumption is wrong bcoz the value of r lies between -1 to $+1$.

So, (i) is the regression line of y on x and (ii) is that of x on y.

Thus, $b_{xy} = -3/2$, and $b_{yx} = -\frac{1}{2}$.

Now, $\frac{b_{yx}}{b_{xy}} = \frac{var(y)}{var(x)}$ $\therefore var(y) = \left(-\frac{1}{2}\right) \times \left(-\frac{3}{2}\right) \times 12 = 4$.

To find $\bar{x}$ and $\bar{y}$, we solve the equations,

$$\begin{array}{ll} (i)\times 2 \Rightarrow & 2x + 4y - 10 = 0 \\ (ii)\times 1 \Rightarrow & 2x + 3y - 8 = 0 \\ & - \quad - \quad + \quad - \\ \hline & y = 2 \text{ and } x = 1 \end{array}$$

$\therefore \bar{x} = 1, \bar{y} = 2$.

Coefficient of Determination :

In case of least square linear regression of y on x, we have seen that

$$r^2 = \frac{var(Y)}{var(y)} = \frac{var(y)-var(e)}{var(y)}$$

$$= 1 - \frac{var(e)}{var(y)}$$

i.e. $r^2 = \frac{s_Y^2}{s_y^2}$ or, $r^2 = 1 - \frac{s_e^2}{s_y^2}$.

where Y_i and e_i are the explained and unexplained parts of y_i, due to the use of regression line corresponding to $x = x_i$.

* Hence, r^2 may be interpreted as the proportion of total variability of y which is explained by it's least square linear regression on x (or, $1-r^2$ may be interpreted as the proportion of total variability of y which is unexplained by it's least square linear regression on x).

* Higher the value of r^2, smaller the value of residual variable and more efficient is the regression equations in predicting values of y.

* The value of r^2 serves as a measure of the worth or usefulness of the linear equation as a predicting formula.

** The quantity r^2 is called the co-efficient of determination. Also $(1-r^2)$ is called the co-efficient of non-determination and it is a measure of deviation from perfect linear relationship.

Ex. 1) If $r_{xy} = 0.3$ and $r_{uv} = 0.6$, does this imply the extent of linear relationship between u and v is twice that between x and y.

Soln. → $r_{xy} = 0.3 \Rightarrow r^2_{xy} = 0.9$ and $r_{uv} = 0.6 \Rightarrow r^2_{uv} = 0.36$.

By defn. of coefficient of determination, we can say that 9% of the total variability is explained by least square linear regression, in case of the variables x and y where as for the variables u and v, 36% of the total variability is explained by the least square linear regression. therefore, the usefulness of linear predicting formula in u and v is fourtimes compare to the variable x and y. Therefore, the extent of linear association between u and v is fourtimes that of x and y.

Therefo re, the extent of linear association between u and v is fourtimes that of x and y.

Ex.2) Why is it necesary to derive two regression line?

Soln.→ If our purpose is to predict y on the basis of fact, then our linear predicting formula is $y = a + bx$, now the line obtained by minimizing the error sum of sauares $\sum_{i=1}^{n}(y_i - a - bx_i)^2$ w.r.to a and b is called the regression line of y on x.

Similarly, to predict x for a given value of y, the predicting formula is $x = c + dy$ and the line obtained by minimising the error sum of sauare $\sum_{i=1}^{n}(x_i - c - dy_i)^2$ w.r.t. c and d is called the regression line of x on y.

Now two regression lines are derived under two different conditions. They should be different. If two variables are exactly linearly related, two regression lines are identical.

***Ex.3)** Consider the array data $\{(x_i, y_{ij}) : j = n_i, i = 1(1)k\}$. Find the least sauares linear regression eauation of y on x.

Soln.→ Note that, the mean of the ith array is $\bar{y}_{x_i} = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$.

Here $\bar{y}_x$ is the regression of y on x. We wish to approximate the regression of y on x by a line, whether the true regression is linear or not. Assuming that → $y = a + bx$, then our model is,

$y_{ij} = (a + bx_i) + e_{ij}$ and $\hat{y}_i = a + bx_i$ is the predicted value of y_{ij}, $j = 1(1)n_i$. when $x = x_i$.

Hence, residual (or error) sum of sauare (RSS or SSE) is

$$S^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}^2 = \sum_i\sum_j (y_{ij} - a - bx_i)^2.$$

To determine 'a' and 'b', we shall use method of least sauares which consists minimizing SSE or RSS with respect to 'a' and 'b'.

Normal eauations are:

$$\frac{\partial S^2}{\partial a} = 0 \quad , \quad \frac{\partial S^2}{\partial b} = 0$$

Hence, $\frac{\partial S^2}{\partial a} = 0 \Rightarrow \sum_i\sum_j (y_{ij} - a - bx_i) = 0$

$$\Rightarrow \sum_i\sum_j y_{ij} = na + b\sum_i n_i x_i$$

$$\Rightarrow \bar{y} = a + b\bar{x} \quad \left[\because \frac{1}{n}\sum_i\sum_j y_{ij} = \bar{y} \text{ and } \frac{1}{n}\sum_i n_i x_i = \bar{x}\right]$$

$$\Rightarrow \boxed{a = \bar{y} - b\bar{x}}$$

and, $\frac{\partial S^2}{\partial b}=0 \Rightarrow \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-a-bx_i)x_i=0$

$\Rightarrow \sum_i\sum_j y_{ij}x_i = a\sum_i n_i x_i + b\sum_i n_i x_i^2$

$\Rightarrow \sum_i n_i x_i \bar{y}_{x_i} = (\bar{y}-b\bar{x})\sum_i n_i x_i + b\sum_i n_i x_i^2$

$\Rightarrow \sum_i n_i x_i \bar{y}_{x_i} - \bar{y}\sum_i n_i x_i = b\left[\sum_i n_i x_i^2 - \bar{x}\sum_i n_i x_i\right]$

$\left[\because \sum_i\sum_j x_i y_{ij} = \sum_i x_i\left(\sum_{j=1}^{n_i} y_{ij}\right) = \sum_i n_i x_i\left(\frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}\right) = \sum_i n_i x_i\cdot\bar{y}_{x_i}\right]$

Dividing both sides by n, we get —

$$\Rightarrow b = \frac{\frac{1}{n}\sum_i n_i x_i\bar{y}_{x_i} - \bar{y}\bar{x}}{\frac{1}{n}\sum_i n_i x_i^2 - \bar{x}^2} = \frac{cov(x,y)}{s_x^2} = r_{xy}\cdot\frac{s_y}{s_x}$$

$$\Rightarrow \boxed{b = r_{xy}\cdot\frac{s_y}{s_x}}$$

Hence, the least sauare linear regression of y on x is

$$y = a+bx$$

$$\therefore y = \bar{y} - b\bar{x} + r_{xy}\frac{s_y}{s_x}$$

$$\therefore \boxed{y = \bar{y} + r_{xy}\frac{s_y}{s_x}(x-\bar{x})}$$

Remark :> For the array data:

Value of x	Values of y	No. of values	Array means
x_1	y_{11} y_{12} ----- y_{1n_1}	n_1	$\bar{y}_{x_1}$
x_2	y_{12} y_{22} ---- y_{2n_2}	n_2	$\bar{y}_{x_2}$
$\vdots$	$\vdots$ $\vdots$ $\vdots$	$\vdots$	$\vdots$
x_k	y_{k1} y_{k2} ---- y_{nn_k}	n_k	$\bar{y}_{n_k}$

Note that, $\bar{x} = \frac{\sum_i n_i x_i}{\sum n_i} = \sum_i n_i x_i/n$ and $\bar{y} = \frac{\sum_i\sum_j y_{ij}}{n} = \frac{\sum_i n_i\bar{y}_i}{n}$

Also, $s_x^2 = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i-\bar{x})^2 = \frac{1}{n}\sum_i n_i x_i^2 - \bar{x}^2$

and, $s_y^2 = \frac{1}{n}\sum_i\sum_j (y_{ij}-\bar{y})^2 = \frac{1}{n}\sum_i\sum_j y_{ij}^2 - \bar{y}^2$

$$\therefore r_{xy} = \frac{cov(x,y)}{s_x\cdot s_y} = \frac{\frac{1}{n}\sum_i\sum_j x_i y_{ij} - \bar{x}\bar{y}}{s_x\cdot s_y} = \frac{\frac{1}{n}\sum_i n_i x_i\bar{y}_{x_i} - \bar{x}\bar{y}}{s_x\cdot s_y}.$$

CORRELATION INDEX :

Correlation ratio measures only the extent of linear relationship between two variables. Sometimes there exists relationship other than linear between two variables which the correlation coefficient fails to capture. In such a situation we look for alternative measures of relationship. For example if there exists a polynomial relationship between two variables x and y, then correlation index can be used to measure that relationship.

▨ Defn. of Correlation Index of order p :

$\boxed{\text{c.v}}$ If y and x are linearly related, so that the regression equation of y on x is $Y = a+bx$, then we have shown that — $r^2 = \frac{Var(Y)}{Var(y)}$.

Following the same line if y can be represented by a p-th degree regression polynomial.

$$Y_p = a_0 + a_1 x + a_2 x^2 + \cdots + a_p x^p.$$

the correlation index of order p is defined by

$$r_p^2 = \frac{Var(Y_p)}{Var(y)}.$$

In fact, r^2 is the proportion of total variability explained by the linear regression equation of y on x and r_p^2 is the proportion of total variability explained by the p-th degree regression equation of y on x. Naturally, $p=1, \Rightarrow r_p^2 = r^2$.

▨ Fitting of p-th degree polynomial regression :

To fit a p-th degree polynomial equation on a set of data $\{(x_i, y_i) : i = 1(1)n\}$, We consider the regression equation

$$Y_{p_i} = a_0 + a_1 x_i + a_2 x_i^2 + \cdots + a_p x_i^p$$ and

estimates the constant a_k's by minimising the sum of square,

$$S^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \cdots - a_p x_i^p)^2$$

Now, $\frac{\partial S^2}{\partial a_K} = 0$ for $K=1(1)p$.

$$\Rightarrow \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i - a_2 x_i^2 - \cdots - a_p x_i^p\right) x_i^K = 0 \quad \forall \ K = 1(1)p.$$

$$\Rightarrow \sum_{i=1}^{n} \left(y_i - Y_{pi}\right) x_i^K = 0 \text{ for } K = 1(1)p.$$

Solving the above set of equations, we can find the valuees of Y_{pi}'s and then calculate the correlation index of order p given by

$$r_p^2 = \frac{\sum_{i=1}^{n} (Y_{pi} - \bar{Y}_p)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Correlation Index (Discussion) : → Suppose we have an observed joint distn. on (x,y) and the problem is to predict y on the basis of x (from the data it is clear enough that x is the cause and y being the corresponding effect i.e. the variable y is only subjected to error).

e.g. a manufacturer of an air conditioning machine wishes to launch a product with a new type of rotary compressure while it is ensured that no. of hours of chilling operation of the machine will be atmost 3 while the outdoor temparature is 112°F. Here the independent variable is the outdoor temparature and dependent one is hours of run.

From the scatterplot of (x,y) if it is evident that the response variable (y) is linearly related with the co-variate (x) i.e. we consider a prediction formula $\psi = a + bx$ in order to regress y, we see that $r^2 = \frac{V(Y)}{V(y)}$, Y being the predicted value.

Clearly $y = Y + e$, where Y being the part of y explained by the least square linear regression of it on x and e being the unexplained part; which is nothing but the residual error. By the normal equation of least square method to predict Y it can be easily stated that $V(y) = V(Y) + V(e)$, thus r^2 is a measure of efficacy of fitting linear regression.

Here r^2, a measure of linear interdependence between x and y is the proportion of total variability in the response variable which can be explained by the linear regression obtained on the basis of co-variate. this idea can easily be generalised to give rise to a similar measure where the response variable is being predicted on the basis of a polynomial regression.

Now, consider the bivariate data $\{(x_i, y_i) : i = 1(1)n\}$. Suppose we wish to approximate the regression equn. of y on x by a p^{th} degree polynomial in x, i.e. assuming

$$Y_p = a_0 + a_1 x + a_2 x^2 + \cdots + a_p x^p \text{, approximately.}$$

Let Y_{p_i} be the predicted value of y_i obtained from its p^{th} degree polynomial regression equation corresponding to $x = x_i$, $i = 1(1)n$. Hence our regression model is $y_i = Y_{p_i} + e_i$, $i = 1(1)n$, where e_i's are the errors in the prediction, the constants $a_0, a_1, \ldots, a_p$ are determined by the method of least squares which consists minimising

$$S^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - Y_{p_i})^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - \cdots - a_p x_i^p)^2$$

w.r.t. $a_0, a_1, \ldots, a_p$.

Normal equations are :

$$\frac{\partial S^2}{\partial a_K} = 0 \quad \cancel{\Rightarrow \sum_{i=1}^{n} \Big(y_i - \sum_{j=0}^{p} a_j x_i^j\Big)}$$

$$\Rightarrow \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - \cdots - a_p x_i^p)\, x_i^K = 0 \quad , K = 1(1)p$$

$$\Rightarrow \sum_{i=1}^{n} e_i x_i^K = 0 \quad , K = 1(1)p.$$

Now, $$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - Y_{p_i} + Y_{p_i} - \bar{y})^2$$
$$= \sum_{i=1}^{n} (y_i - Y_{p_i})^2 + \sum_{i=1}^{n} (Y_{p_i} - \bar{y})^2 \quad \text{——} ①$$

[∵ product term vanishes due to normal equation]

Also, by first normal equation:

$$\sum_{i=1}^{n} e_i = 0 \Rightarrow \sum (y_i - Y_{p_i}) = 0$$
$$\Rightarrow \bar{y} = \bar{Y}_p .$$

Putting $\bar{y} = \bar{Y}_p$ in ①, we get → $\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - Y_{p_i})^2 + \sum (Y_{p_i} - \bar{Y}_p)^2$

$$\Rightarrow V(y) = V(e) + V(Y_p) \quad \text{——} ②$$

⇒ Total variability of y = Unexplained or error variability + Explained variability — due to the use of the p^{th} degree polynomial regression as a predicting formula.

The smaller the unexplained variability in comparison with the total variability of y, the better the predicting formula. Hence, as a measure of usefulness or efficiency of the p^{th} degree least squares polynomial regression, We define, $r_p^2 = 1 - \dfrac{\text{Unexplained variability by the pth degree polynomial regressions}}{\text{Total variability of } y}$

$$= 1 - \frac{\sum_{i=1}^{n}(y_i - Y_{pi})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$= \frac{\sum_i (Y_{pi} - \bar{Y}_p)^2}{\sum_i (y_i - \bar{y})^2} \qquad [\text{Using } ②]$$

$$= \frac{var(Y_p)}{var(y)}$$

The positive square root of r_p^2 is called the correlation index and we define $r_p = \sqrt{\dfrac{var(Y_p)}{var(y)}}$.

Note that, r_p^2 is the proportion of total variability explained by the pth degree polynomial regression of y on x.

▨ PROPERTIES :→

① . $0 \le r_p \le 1$. When does equality hold?

Proof :→ We have ⟶ $\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - Y_{pi})^2 + \sum_i (Y_{pi} - \bar{Y}_p)^2$

$$\Rightarrow \sum_i (y_i - \bar{y})^2 \geq \sum_i (Y_{pi} - \bar{Y}_p)^2 \quad [\because \bar{y} = \bar{Y}_p]$$

$$\therefore r_p^2 = \frac{\sum_i (Y_{pi} - \bar{Y}_p)^2}{\sum_i (y_i - \bar{y})^2} \le 1$$

$$\therefore 0 \le r_p \le 1.$$

Equality cases :→ i) $r_p = 0 \Rightarrow \sum_i (Y_{pi} - \bar{Y}_p)^2 = 0$

$$\Rightarrow Y_{pi} = \bar{Y}_p \quad \forall\ i = 1(1)n,$$

⇒ the predicting formula gives a constant value, ∀ i.

⇒ there is no use of pth degree polynomial equations in predicting the y values.

ii) $r_p = 1$

iff $\sum_{i=1}^{n}(y_i - Y_{pi})^2 = 0$

iff $y_i = Y_{pi} = a_0 + a_1 x_i + \cdots + a_p x_i^p \quad \forall\ i$

iff y is an exact polynomial of degree p as far as the given data is concerned.

$\boxed{\text{C.U.}}$

② . Correlation Index is an increasing functions of its degree p. i.e. $r_p^2 \geq r_{p-1}^2$. Then show equality case.

Proof :→ Let the pth degree and the (p-1)th degree regression equation can be represented as ——

$$Y_{p_i} = a_0 + a_1x_i + a_2x_i^2 + \cdots + a_px_i^p \text{ and}$$

$$Y_{\overline{p-1}\,i} = a_0' + a_1'x_i + a_2'x_i^2 + \cdots + a_{p-1}'x_i^{p-1}.$$

Now,
$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(y_i - Y_{p_i} + Y_{p_i} - \bar{y})^2$$

$$= \sum_{i=1}^{n}(y_i - Y_{p_i})^2 + \sum_{i=1}^{n}(Y_{p_i} - \bar{y})^2$$

[∵ Cross product term vanishes due to normal equation]

$$= \sum_{i=1}^{n}(y_i - Y_{p_i})^2 + \sum_{i=1}^{n}(Y_{p_i} - \bar{Y}_p)^2 \quad [\bar{Y}_p = \bar{y} \text{ by 1st normal eqn.}]$$

$\longrightarrow$ ①

Product term :→

$$\sum_{i=1}^{n}(y_i - Y_{p_i})(Y_{p_i} - \bar{y})$$

$$= \sum_{i=1}^{n}(y_i - a_0 - a_1x_i - a_2x_i^2 - \cdots - a_px_i^p)Y_{p_i} - \bar{y}\sum_{i=1}^{n}(y_i - a_0 - a_1x_i - \cdots - a_px_i^p)$$

$$= \sum_{i=1}^{n}(y_i - a_0 - a_1x_i - a_2x_i^2 - \cdots - a_px_i^p)\left(\sum_{k=0}^{p}a_kx_i^k\right) - 0$$ [By 1st normal eqn., the 2nd part is 0]

$$= \sum_{k=0}^{p}a_k\underbrace{\sum_{i=1}^{n}(y_i - a_0 - a_1x_i - a_2x_i^2 - \cdots - a_px_i^p)x_i^k}_{=0}$$

$= 0$ [from the kth normal equation]

↑ (NOT FOR EXAM, FOR CONCEPT)

$y_i = Y_{p_i} + e_{p_i}$

$\sum e_{p_i} = \sum_i (y_i - Y_{p_i})$

$= 0$

From the 1st normal equation.

Since $Y_{p_i} = a_0 + a_1x_i + a_2x_i^2 + \cdots + a_px_i^p$, the constants have been obtained by the least square method.

By defn. of the least square method, then

$$\sum_{i=1}^{n}(y_i - a_0 - a_1x_i - a_2x_i^2 - \cdots - a_px_i^p)^2 \leq \sum_{i=1}^{n}(y_i - b_0 - b_1x_i - \cdots - b_px_i^p)^2$$

whatever the alternative set of constants $b_0, b_1, \ldots, b_p$ may be.

Let us take, $b_0 = a_0'$, $b_1 = a_1'$, ..., $b_{p-1} = a_{p-1}'$, $b_p = 0$, we have

then $$\sum_{i=1}^{n}(y_i - a_0 - a_1x_i - \cdots - a_px_i^p)^2 \leq \sum_{i=1}^{n}(y_i - a_0' - a_1'x_i - \cdots - a_{p-1}'x_i^{p-1})^2$$

i.e. $$\sum_{i=1}^{n}(y_i - Y_{p_i})^2 \leq \sum_{i=1}^{n}(y_i - Y_{\overline{p-1}\,i})^2 \longrightarrow (*)$$

From ① → $\sum_{i=1}^{n} (y - Y_{pi})^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \sum_{i=1}^{n} (Y_{pi} - \bar{Y}_p)^2$ —— ②

Putting $p = \overline{p-1}$ in equn. ②.

$$\sum_{i=1}^{n} (y - Y_{\overline{p-1}i})^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \sum_{i=1}^{n} (Y_{\overline{p-1}i} - \bar{Y}_{\overline{p-1}})^2 \text{ —— ③}$$

Putting ② and ③ in the inequality ⊛, we get —

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 - \left(\sum_{i=1}^{n} (Y_{pi} - \bar{Y}_p)^2 \le \sum_{i=1}^{n} (y_i - \bar{y})^2 - \sum_{i=1}^{n} (Y_{\overline{p-1}i} - \bar{Y}_{p-1})^2\right.$$

$$\Rightarrow \sum_{i=1}^{n} (Y_{pi} - \bar{Y}_p)^2 \geq \sum_{i=1}^{n} (Y_{\overline{p-1}i} - \bar{Y}_{p-1})^2$$

$$\Rightarrow \frac{\sum_{i=1}^{n} (Y_{pi} - \bar{Y}_p)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \geq \frac{\sum_{i=1}^{n} (Y_{\overline{p-1}i} - \bar{Y}_{p-1})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

$\Rightarrow r_p^2 \geq r_{p-1}^2$ for $p = 2, 3, \dots, n-1$.

i.e. $r_1^2 \le r_2^2 \le r_3^2 \le \dots \le r_{p-1}^2 \le r_p^2 \le r_{p+1}^2 \le \dots \le r_{n-1}^2$

Equality Case :→ i.e. $\boxed{r_p^2 = r_{p-1}^2}$

$$\Rightarrow \sum_{i=1}^{n} (y_i - Y_{pi})^2 = \sum_{i=1}^{n} (y_i - Y_{\overline{p-1}i})^2$$

$\Rightarrow$ if $Y_{pi} = Y_{\overline{p-1}i}$ ∀ i.

⇒ the pth and (p-1)th degree polynomial regressions are identical, means pth degree polynomial can be fitted as (p-1)th degree fitting.

i.e. there is no use of pth degree polynomial regression in stead of (p-1)th degree polynomial regression.

C.U.

Interpretation of the property : $r_p^2 \geq r_{p-1}^2$.

The property states that r_p^2 is a non-decreasing function of the degree (p) at the polynomial used in prediction purpose. therefore, if we use higher degree polynomial as a predicting formula, we will have less error variability, or, another way —

"The more is the degree of the polynomial, the more proportion of variability in the response variable can expected to be explained".

Also note that the pth degree polynomial equatiom —

$$Y_p = a_0 + a_1x + a_2x^2 + \dots + a_{p-1}x^{p-1} + a_px^p.$$

reduces to (p-1)th degree polynomial regression equation, where $a_p = 0$.

Therefore, it is logically clear that $r_p^2 \geq r_{p-1}^2$.

Correlation Ratio :-

Here we shall introduce a more general measure of the extent of dependence of one variable on the other. Let n pairs of values $(x_i, y_i) : i = 1(1)n$ are arranged in arrays of y according to fixed values of y.

Value of x	Value of y	Means
x_1	$y_{11}\ y_{12} \cdots\cdots y_{1n_1}$	$\overline{y_1}$
x_2	$y_{21}\ y_{22} \cdots\cdots y_{2n_2}$	$\overline{y_2}$
$\vdots$	$\vdots \quad \vdots \qquad \vdots$	$\vdots$
x_k	$y_{k1}\ y_{k2} \cdots\cdots y_{kn_k}$	$\overline{y_k}$

Here, $\overline{y_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$, is the mean of the ith array.

Note that, $\overline{y} = \dfrac{\sum_{i=1}^{k} n_i\overline{y_i}}{\sum_{i=1}^{k} n_i}$

For fixed x, the array mean is $\overline{y_x}$, then the array mean $\overline{y_x}$ when x is given, as a function of x, is called the regression equation of y on x.

It has been shown that the regression eqn. is the best predicting formula in terms of minimum SSE. If we use the true regression equation as the predicting formula, our model is then $y_{ij} = \overline{y_i} + e_{ij}$, $i = 1(1)k$, $j = 1(1)n_i$.

Note that — $\sum_i\sum_j (y_{ij} - \overline{y})^2$

$$= \sum_i\sum_j (y_{ij} - \overline{y_i} + \overline{y_i} - \overline{y})^2$$

$$= \sum_i\sum_j (y_{ij} - \overline{y_i})^2 + \sum_i\sum_j (\overline{y_i} - \overline{y})^2$$

[product term vanishes since $\sum_j (y_{ij} - \overline{y_i}) = \sum_j e_{ij} = 0$]

$$\geq \sum_i\sum_j (y_{ij} - \overline{y_i})^2$$

with equality iff $\overline{y} = \overline{y_i}$

As our model is $y_{ij} = \bar{y}_i + e_{ij}$, then SSE or unexplained variability due to use of regression equation of y on x as a predicting formula for y is

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$$

We have →
$$\sum_i\sum_j(y_{ij}-\bar{y})^2 = \underbrace{\sum_i\sum_j(y_{ij}-\bar{y}_i)^2}_{\text{(within array variability)}} + \underbrace{\sum_i n_i(\bar{y}_i-\bar{y})^2}_{\text{(between array variability)}} \quad \text{——} (1)$$

⇒ Total variability of y = Unexplained variability + explained variability

The smaller the unexplained variability $\sum_i\sum_j(y_{ij}-\bar{y}_i)^2$ compare to the total variability $\sum_i\sum_j(y_{ij}-\bar{y})^2$, the better the predicting formula. Hence, as a measure of usefulness or efficacy of the regression eqn. of y on x as a predicting formula for y values, we define,

$$e^2_{yx} = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2}$$

$$\text{i.e. } e^2_{yx} = \frac{\sum_{i=1}^{K} n_i(\bar{y}_i-\bar{y})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2} \qquad [\text{by } (1)]$$

The positive square root of (e^2_{yx}), i.e.

$$e_{yx} = +\sqrt{\frac{\sum_{i=1}^{k} n_i(\bar{y}_i-\bar{y})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2}}$$

is called Correlation ratio. 10.9.

▨ When two variables x and y are related among themselves such that their relationship can't be explained by any polynomial function, then to ascertain the degree of relationship between them we need a third kind of measure called Regression Ratio.

PROPERTIES :>

1) $0 \le e_{yx} \le 1$. Interpret the equality cases or extreme cases.

Proof :>

For n pairs of values (x_i, y_i), $i = 1(1)n$ which are arranged in arrays, let the mean of the i th array is $\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$, $i = 1(1)k$.

We know, $y_{ij} = \bar{y}_i + e_{ij}$; $i = 1(1)k$, $j = 1(1)n_i$.

Note that —

$$\sum_i \sum_j (y_{ij} - \bar{y})^2$$

$$= \sum_i \sum_j (y_{ij} - \bar{y}_i + \bar{y}_i - \bar{y})^2$$

$$= \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 + \sum_i \sum_j (\bar{y}_i - \bar{y})^2$$

[product term vanishes as $\sum_j (y_{ij} - \bar{y}_i) = \sum_j e_{ij} = 0$]

$$\Rightarrow \sum_i \sum_j (y_{ij} - \bar{y})^2 = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 + \sum_i n_i (\bar{y}_i - \bar{y})^2$$

$\Rightarrow$ [redacted] $\sum_i n_i (\bar{y}_i - \bar{y})^2 \le \sum_i \sum_j (y_{ij} - \bar{y})^2$

[$\because \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 \geq 0$]

$$\Rightarrow \frac{\sum_i n_i (\bar{y}_i - \bar{y})^2}{\sum_i \sum_j (y_{ij} - \bar{y})^2} \le 1$$

$$\Rightarrow 0 \le e_{yx}^2 \le 1.$$

Equality Cases :>

$\boxed{c.u}$ i) $e_{yx}^2 = 0$

$\Rightarrow \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2 = 0$

$\Rightarrow \bar{y}_i = \bar{y} \quad \forall i.$

$\Rightarrow$ the regression of y on x is constant, i.e., the regression of y on x has no use as a predicting formula for y values.

ii) $\underline{e^2_{yx} = 1}$

$$\Rightarrow \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 = 0$$

$$\Rightarrow y_{ij} = \bar{y}_i \quad \forall \; i, j.$$

i.e., $y = \bar{y}_x$, a function of x.

$\Rightarrow y$ can be expressed as an exact function of x.

i.e. there is an exact functional relationship such as $y = f(x)$.

$\boxed{\text{C.V.}}$

2) $r^2 \leq e^2_{yx}$. Explain the equality cases.

Proof: → Consider the array data $\{(x_i, y_{ij}) : i = 1(1)k, j = 1(1)n_i\}$.
Let $Y_i = a + bx_i$ be the fitted regression line of y on x.

Note that →
$$\sum_i \sum_j (y_{ij} - Y_i)^2 = \sum_i \sum_j (y_{ij} - \bar{y}_i + \bar{y}_i - Y_i)^2$$
$$= \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 + \sum_i n_i (\bar{y}_i - Y_i)^2 \quad \text{——} \; (1)$$

[the cross product term vanishes as $\sum_j (y_{ij} - \bar{y}_i) = 0$]

Now, we consider, $r^2 = 1 - \frac{var(e)}{var(y)} = 1 - \frac{\sum_i \sum_j (y_{ij} - Y_i)^2}{\sum_i \sum_j (y_{ij} - \bar{y})^2}$

$$\Rightarrow (1 - r^2) = \frac{\sum_i \sum_j (y_{ij} - Y_i)^2}{\sum_i \sum_j (y_{ij} - \bar{y})^2} \Rightarrow \sum_i \sum_j (y_{ij} - Y_i)^2 = \left\{\sum_i \sum_j (y_{ij} - \bar{y})^2\right\}(1 - r^2) \quad \text{——} (*)$$

and, $$(1 - e^2_{yx}) = \frac{\sum_i \sum_j (y_{ij} - \bar{y}_i)^2}{\sum_i \sum_j (y_{ij} - \bar{y})^2} \Rightarrow \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 = \left\{\sum_i \sum_j (y_{ij} - \bar{y})^2\right\}(1 - e^2_{yx}) \quad \text{——} (**)$$

Putting $(*)$ and $(**)$ in (1), we get —

$$\left\{\sum_i \sum_j (y_{ij} - \bar{y})^2\right\}\left\{1 - r^2\right\} = \left\{\sum_i \sum_j (y_{ij} - \bar{y})^2\right\}\left\{1 - e^2_{yx}\right\} + \sum_i n_i (\bar{y}_i - Y_i)^2$$

$$\Rightarrow \sum_i n_i (\bar{y}_i - Y_i)^2 = \left\{\sum_i \sum_j (y_{ij} - \bar{y})^2\right\}\left\{e^2_{yx} - r^2\right\}$$

$$\Rightarrow e^2_{yx} - r^2 = \frac{\sum_i n_i (\bar{y}_i - Y_i)^2}{\sum_i \sum_j (y_{ij} - \bar{y})^2} \geq 0$$

$$\Rightarrow e^2_{yx} \geq r^2$$

Equality cases: →
$e^2_{yx} = r^2 \Rightarrow \sum_{i=1}^{k} n_i (\bar{y}_i - Y_i)^2 = 0$, i.e. if $Y_i = \bar{y}_i$
$\forall \; i = 1(1)k$.

$\Rightarrow$ the regression equation of y on x is linear.

i.e. if the regression lines of y on x passes through the array mean.

▨ **Remark:-** $\boxed{\text{C.U.}}$

1) Show that $0 \le d^2_{yx} \le 1$, where $d^2_{yx} = e^2_{yx} - r^2$. Explain the equality cases.

Proof:→ We know $r^2 \le e^2_{yx}$, we have now

$$e^2_{yx} - r^2 = \frac{\sum_{i=1}^{k} n_i(\bar{y}_i-\bar{y})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2} - \frac{\sum_{i=1}^{k} n_i(Y_i-\bar{y})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2} \longrightarrow ①$$

where $Y_i = a+bx_i$, and,

$$\sum_{i=1}^{k} n_i(\bar{y}_i-\bar{y})^2$$

$$= \sum_{i=1}^{k} n_i\{(\bar{y}_i - Y_i) + (Y_i-\bar{y})\}^2$$

$$= \sum_i n_i(\bar{y}_i - Y_i)^2 + \sum_i n_i(Y_i-\bar{y})^2$$

[the cross product term vanishe due to first normal equation, i.e $\sum n_i(\bar{y}_i - Y_i) = 0$]

i.e. $\sum_{i=1}^{k} n_i(\bar{y}_i-\bar{y})^2 - \sum_{i=1}^{k} n_i(Y_i-\bar{y})^2_i = \sum_{i=1}^{k} n_i(\bar{y}_i - Y_i)^2$

So, from ①, $e^2_{yx} - r^2 = \dfrac{\sum_{i=1}^{k} n_i(\bar{y}_i - Y_i)^2}{\sum_i\sum_j (y_{ij}-\bar{y})^2}$

Clearly, $e^2_{yx} - r^2 = d^2_{yx} \geqslant 0$.

Also, we have $\longrightarrow$ $r^2 \le e^2_{yx} \le 1$ and $0 \le r^2 \le 1$

Now, $d^2_{yx} = e^2_{yx} - r^2$ will be maximum when e^2_{yx} is maximum and r^2 is minimum, i.e., when $e^2_{yx} = 1$ and $r^2 = 0$. Hence, $d^2_{yx} \le 1$.

Therefore, $0 \le d^2_{yx} \le 1$.

Equality cases :→

1) $d^2_{yx} = 0 \Rightarrow e^2_{yx} = r^2 \Rightarrow \sum_{i=1}^{n} n_i(\overline{y_i} - Y_i)^2$

$\Rightarrow \overline{y_i} = Y_i = a + bx_i \ , i = 1(1)k.$

⇒ the regression equation of y on x is linear.

2) $d^2_{yx} = 1$ iff $e^2_{yx} = 1$ and $r^2 = 0$

iff y is an exact function of x and there is no linear relationship i.e. y is an exact non-linear function of x.

What does $d^2_{yx} = e^2_{yx} - r^2_{yx}$ measure ? / Significance of d^2_{yx}.

Ans:→ Significance of d^2_{yx} :→

$$d^2_{yx} = e^2_{yx} - r^2 = \frac{\sum_{i=1}^{k} n_i(\overline{y_i} - Y_i)^2}{\sum_i \sum_j (y_{ij} - \overline{y})^2}$$

Note that $d^2_{yx} = 0 \Rightarrow \overline{y_i} = Y_i = a + bx_i \ , i = 1(1)k.$

i.e. the regression equation of y on x is linear.

If $\overline{y_i} \neq a + bx_i$, then $d^2_{yx} = e^2_{yx} - r^2 > 0$.

The larger the deviations $(\overline{y_i} - a - bx_i)$, the higher the value of d^2_{yx}. Hence, the difference $d^2_{yx} = (e^2_{yx} - r^2)$ measure the extent to which the regression equation of y on x departs from linearity.

C.U.

Remark 2. Note that, a measure of usefulness or efficacy of the true regression equation of x on y as a predicting formula is

$$e^2_{xy} = \frac{\sum_{i=1}^{k} n_i(\overline{x_i} - c - dy_i)^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x})^2} \quad , \text{where } c + dy_i = X_i.$$

In general, $e^2_{yx} \neq e^2_{xy}$, i.e., correlation ratio is not symmetric. That is due to the fact that regression equation of y on x, that is $\overline{y}_x$ and regression eqn. of x on y $\overline{x}_y$ are not identical in general.

Now, $e^2_{yx} = e^2_{xy} = 1$, iff y is an exact function of x, say, $y = f(x)$ and x is an exact function of y, say $x = g(y)$.

iff the variables x and y are exactly functionally related and the functional relation is one-to-one, i.e its inverse exists.

$\boxed{\text{C.U.}}$

PROPERTY 3) → The correlation ratio e_{yx} is the product moment correlation coefficient between y and the array mean of y corresponding to x. 10.11

Proof :→ The correlation coefficient between y and $\bar{y}_x$ is

$$r_{y\bar{y}_x} = \frac{\text{cov}(y, \bar{y}_x)}{\sqrt{\text{var}(y)}\sqrt{\text{var}(\bar{y}_x)}}$$

$$= \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})(\bar{y}_i-\bar{y})}{\sqrt{\sum_i\sum_j(y_{ij}-\bar{y})^2}\sqrt{\sum_{i=1}^{k}n_i(\bar{y}_i-\bar{y})^2}}$$

$$= \frac{\sum_{i=1}^{k}(\bar{y}_i-\bar{y})\sum_{j=1}^{n_i}(y_{ij}-\bar{y})}{\sqrt{\sum_i\sum_j(y_{ij}-\bar{y})^2}\sqrt{\sum_i n_i(\bar{y}_i-\bar{y})^2}}$$

$$= \frac{\sum_{i=1}^{k}n_i(\bar{y}_i-\bar{y})^2}{\sqrt{\sum_i\sum_j(y_{ij}-\bar{y})^2}\sqrt{\sum_i n_i(\bar{y}_i-\bar{y})^2}} \quad \left[\because \sum_{j=1}^{n_i}(y_{ij}-\bar{y}) = n_i(\bar{y}_i-\bar{y})\right]$$

$$= \sqrt{\frac{\sum_{i=1}^{k}n_i(y_i-\bar{y})^2}{\sum_i\sum_j(y_{ij}-\bar{y})^2}}$$

$$= e_{yx}.$$

$\boxed{\text{C.U.}}$

PROPERTY 4) $0 \le r^2 \le r_p^2 \le e^2_{yx} \le 1$. Explain boundary cases. 10.9.

Proof :→

Note that →

$$\sum_i\sum_j(y_{ij}-\bar{y})^2 = \sum_i\sum_j(y_{ij}-\bar{y}_i+\bar{y}_i-\bar{y})^2$$

$$= \sum_i\sum_j(y_{ij}-\bar{y}_i)^2 + \sum_i n_i(\bar{y}_i-\bar{y})^2 \quad \text{——} (1)$$

[product term vanishes as $\sum_j(y_{ij}-\bar{y}_i) = 0$]

$$\sum_i n_i(\bar{y}_i - \bar{y})^2 = \sum_i n_i(\bar{y}_i - Y_{pi})^2 + \sum_i \overset{n_i}{\wedge}(Y_{pi} - \bar{y})^2 \quad \text{——} \ ②$$

Product term vanishes since

$$2\sum_i n_i \underbrace{(\bar{y}_i - Y_{pi})}_{e_{pi}} (Y_{pi} - \bar{y})$$

$$= 2\Big[\sum_i n_i e_{pi} Y_{pi} - \bar{y} \underbrace{\sum_i n_i e_{pi}}_{\overset{\shortparallel}{0}}\Big]$$

$$= 2\Big[a_0 \sum_i n_i e_{pi} + a_1 \sum_i n_i e_{pi} x_i + \cdots\cdots\Big]$$

$$= 0$$

as the normal equation is obtained by minimizing $\sum_i n_i(\bar{y}_i - a_0 x - a_1 x_i - \cdots\cdots - a_p x_i^p)^2$ w.r.to a_i's.

$$① \Rightarrow \frac{\sum_i n_i(\bar{y}_i - \bar{y})^2}{\sum_i \sum_j (y_{ij} - \bar{y})^2} \le 1, \quad \Big[\because \sum_i \sum_j (y_{ij} - \bar{y})^2 \geq \sum_i n_i(\bar{y}_i - \bar{y})^2\Big]$$

i.e. $e^2_{yx} \le 1$. ——— (*)

$$② \Rightarrow \sum_i n_i(\bar{y}_i - \bar{y})^2 \geq \sum_i n_i(Y_{pi} - \bar{y})^2$$

$$\Rightarrow \sum_i n_i(\bar{y}_i - \bar{y})^2 \geq \sum_i n_i(Y_{pi} - \bar{Y}_p)^2, \text{ since } \bar{Y}_p = \bar{y}.$$

Dividing both sides by $s_y^2 = \sum_i \sum_j (y_{ij} - \bar{y})^2$

i.e. $$\frac{\sum_i n_i(\bar{y}_i - \bar{y})^2}{\sum_i \sum_j (y_{ij} - \bar{y})^2} \geq \frac{\sum_i n_i(Y_{pi} - \bar{Y}_p)^2}{\sum_i \sum_j (y_{ij} - \bar{y})^2}$$

$\Rightarrow e^2_{yx} \geq r_p^2 \ \forall\ p = 1(1)n-1$. ——— (**)

Obviously, we have → $0 \le r^2 \le r_2^2 \le \cdots\cdots \le r_p^2 \le \cdots \le r_{n-1}^2$ ——— (***)

By (*), (**), (***), we have

$$0 \le r^2 \le r_2^2 \le \cdots\cdots \le r_p^2 \le e^2_{yx} \le 1, \ (\underline{\text{Proved}})$$

FURTHER EXAMPLES :→

Ex. 1. If two pairs of values (x_1, y_1) and (x_2, y_2) are observed on two variables, what will be the correlation coefficient?

Soln.→ Given data : (x_1, y_1) and (x_2, y_2)

The least square linear regression eqn. of y on x is the line $Y = a+bx$ for which $\sum_{i=1}^{2} e_i^2 = \sum_{i=1}^{2}(y_i - a - bx_i)^2$ is minimum. the line passes through the points (x_1, y_1) and (x_2, y_2) gives $e_1 = 0 = e_2$, i.e. minimizes $\sum_{i=1}^{2} e_i^2$. Hence,

$y_i = Y_i + e_i \Rightarrow y_i = Y_i$ as $e_i = 0$, for $i = 1, 2$.

$\Rightarrow y_i = a + bx_i$ for $i = 1, 2$.

Hence, the relationship is exactly linear as far as the given data is concerned.

Hence, $r = \pm 1$ and $r = \begin{cases} +1 & \text{if } b > 0 \\ -1 & \text{if } b < 0 \end{cases}$

Ex. 2. Show that, if x' and y' are the deviations of the variables x and y from their mean, then

$$r = 1 - \frac{1}{2n}\sum_{i=1}^{n}\left(\frac{x_i'}{s_x} - \frac{y_i'}{s_y}\right)^2$$

$$= -1 + \frac{1}{2n}\sum_{i=1}^{n}\left(\frac{x_i'}{s_x} + \frac{y_i'}{s_y}\right)^2$$

Hence deduce that → $-1 \leq r \leq 1$.

Soln.→ Let $\{(x_i, y_i) : i = 1(1)n\}$ be the given data.

Here, $x_i' = x_i - \bar{x}$, $y_i' = y_i - \bar{y}$.

Now, $1 - \frac{1}{2n}\sum_{i=1}^{n}\left(\frac{x_i'}{s_x} - \frac{y_i'}{s_y}\right)^2$

$$= 1 - \frac{1}{2n}\sum_{i=1}^{n}\left(\frac{x_i'^2}{s_x^2} + \frac{y_i'^2}{s_y^2} - \frac{2x_i'y_i'}{s_xs_y}\right)$$

$$= 1 - \frac{1}{2\cancel{n}}\cdot\frac{\sum_i(x_i - \bar{x})^2}{\frac{1}{\cancel{n}}\sum_i(x_i - \bar{x})^2} - \frac{1}{2\cancel{n}}\cdot\frac{\sum_i(y_i - \bar{y})^2}{\frac{1}{\cancel{n}}\sum_i(y_i - \bar{y})^2} + \frac{2}{2n}\cdot\sum_i\frac{x_i'y_i'}{s_xs_y}$$

$$= 1 - \frac{1}{2} - \frac{1}{2} + \frac{1}{n}\frac{\sum_i(x_i - \bar{x})(y_i - \bar{y})}{s_xs_y} = \frac{\frac{1}{n}\sum_i(x_i - \bar{x})(y_i - \bar{y})}{s_xs_y} = r$$

(proved)

Similarly, $-1+\frac{1}{2n}\sum_{i=1}^{n}\left(\frac{x_i'}{s_x}+\frac{y_i'}{s_y}\right)^2$

$$= -1+\frac{1}{2}+\frac{1}{2}+\frac{1}{n}\sum_i \frac{x_i'y_i'}{s_x s_y}$$

$$= \frac{\frac{1}{n}\sum_i (x_i-\bar{x})(y_i-\bar{y})}{s_x s_y}$$

$= r$. (proved)

∴ As, $\frac{1}{2n}\sum_i\left(\frac{x_i'}{s_x}-\frac{y_i'}{s_y}\right)^2$ and $\frac{1}{2n}\sum_i\left(\frac{x_i'}{s_x}+\frac{y_i'}{s_y}\right)^2$ are ≥ 0

so, $1-r \geq 0$ and $1+r \geq 0$.

$\Rightarrow -1 \leq r \leq 1$. (Proved)

Ex. 3. If θ is the angle between two regression lines, then show that → $\sin\theta \leq 1-r^2$.

Soln. → we have → $\tan\theta = \frac{1-r^2}{|r|}\cdot\frac{s_x s_y}{s_x^2+s_y^2}$ ——— (1)

Now, $\frac{s_x^2+s_y^2}{2} \geq \sqrt{s_x^2 s_y^2}$ $\left[\because (s_x-s_y)^2 \geq 0 \Rightarrow s_x^2+s_y^2-2s_xs_y \geq 0 \Rightarrow \frac{s_x^2+s_y^2}{2} \geq s_xs_y\right]$

$$\Rightarrow \frac{s_x s_y}{s_x^2+s_y^2} \leq \frac{1}{2}$$

From (1), we get

$$\therefore \tan\theta \leq \frac{1-r^2}{2|r|}$$

$$\Rightarrow \frac{\sin^2\theta}{\cos^2\theta} \leq \frac{(1-r^2)^2}{4r^2}$$

$$\Rightarrow \frac{\sin^2\theta}{\sin^2\theta+\cos^2\theta} \leq \frac{(1-r^2)^2}{4r^2+(1-r^2)^2}$$

$$\Rightarrow \sin^2\theta \leq \frac{(1-r^2)^2}{(1+r^2)^2}$$

$$\Rightarrow \sin\theta \leq \frac{1-r^2}{1+r^2} \leq 1-r^2$$

C.U.

IMPORTANT QUESTIONS :→

1> What is the implication of the measure $e^2_{yx} - r^2$?

Ans:→ The relation $e^2_{yx} - r^2$ measures the departure of the regression equation from linearity.

2> Interprete the following cases :—

(i) $r^2 = e^2_{yx}$ but $e^2_{xy} \neq r^2$. 10.9.

⇒ "the regression equation of y on x is linear but that of x on y is non-linear."

i.e. the regression of y on x is linear but y and x are not truely or fractionally related.

✓ (ii) $r^2 = e^2_{yx} = 1$. 10.9.

⇒ y is truely a function of x.

(iii) $e^2_{yx} = e^2_{xy} = 1$. 10.9.

⇒ x and y are truely or fractionally related, i.e., if $y = g(x)$ then $x = g^{-1}(y)$.

(iv) $e^2_{yx} = e^2_{xy}$.

⇒ the relation gives unique y for a given x and unique x for a given y.

(v) $e^2_{yx} = e^2_{xy} = r^2$. 10.9.

⇒ the regression equation of y on x and the regression equation of x on y are both linear.

(vi) $r^2_p - r^2 = 1$.

⇔ $r^2_p = 1$ but $r^2 = 0$.

⇔ y is an exact polynomial of degrees of p but it is not linear.

⇔ y is a non-linear polynomial of degree $p \geq 2$.

* $r^2 = e^2_{yx}$ ⇒ exact linear relationship between x and y.

(vii) $e^2_{xy} = e^2_{yx} < 1$.

⇒ there is no exact functional relationship between x and y but the regression of y on x as a predicting formula for y values and the regression of x on y as a predicting formula for x values are equally efficient.

(viii) $r^2 < r_2^2 < r_3^2 = e^2_{yx} = 1$.

⇒ Total variability in the response variable, y can be explained on the basis of co-variate x.

(ix) $r^2 < r_2^2 < r_3^2 = e^2_{yx} = 0.5$.

⇒ there is not any perfect linear relationship between x and y. y can be expressed truely as a polynomial of degree 3 on the basis of x, and by the function $y = f(x)$, $f(x)$ be a polynomial of degree 3, we can explain half of the total variability in the response variable. y be the polynomial regression of order 3 obtained on the basis of x.

3) i) What is measured by $(r_p^2 - r^2)$?

ii) Interprete the case : $e^2_{yx} = r_p^2$. What is measured by $(e^2_{yx} - r_p^2)$?

3) Let x and y be subject to observational errors, so that what one observes really are $x' = x + \varepsilon_x$ and $y' = y + \varepsilon_y$ instead of x and y. If ε_x and ε_y are independent of x and y, and if ε_x and ε_y are also mutually independent, show that the correlation coefficient between x and y becomes numerically smaller owing to the errors. (This is called the 'attenuation effect') <u>10.16</u>

<u>Soln.</u> → Here the variables $(x, y, \varepsilon_x, \varepsilon_y)$ are such that each pair in the set except x, y is uncorrelated.

$$r_{xy} = r \text{ and } r_{x'y'} = r'.$$

$$r' = \frac{\text{cov}(x', y')}{\sqrt{\text{var}(x')}\sqrt{\text{var}(y')}}$$

$$\therefore \text{cov}(x', y') = \text{cov}(x + \varepsilon_x, y + \varepsilon_y)$$

$$= \text{cov}(x, y) + \text{cov}(x, \varepsilon_y) + \text{cov}(\varepsilon_x, y) + \text{cov}(\varepsilon_x + \varepsilon_y)$$

$$= \text{cov}(x, y)$$

$$\therefore \text{var}(x') = \text{var}(x + \varepsilon_x)$$
$$= \text{var}(x) + \text{var}(\varepsilon_x)$$

Similarly, $\text{var}(y') = \text{var}(y + \varepsilon_y)$
$$= \text{var}(y) + \text{var}(\varepsilon_y)$$

Now, $$r'^2 = \frac{\text{cov}^2(x, y)}{\{V(x) + V(\varepsilon_x)\}\{V(y) + V(\varepsilon_y)\}}$$

Now, $V(x) + V(\varepsilon_x) \geq V(x)$

and $V(y) + V(\varepsilon_y) \geq V(y)$

i.e. $$\frac{1}{V(x) + V(\varepsilon_x)} \cdot \frac{1}{V(y) + V(\varepsilon_y)} \leq \frac{1}{V(x)\,V(y)}$$

$\therefore \quad (r')^2 \leq r^2$

i.e. $|r'| \leq |r|$

CHAPTER – 6

RANK CORRELATION

C.U.

Rank Correlation Coefficient :—

For calculating the product-moment correlation coefficient, it is essential that measurements on the two characters are available. But in many cases, the characters may not be numerically measurable or, even if it is numerically measurable, the data is not given numerically. Then the product moment correlation coefficient can not be calculated for the data. Suppose that it is possible to arrange the individuals or items according to the degree to which they possess a character under enquiry. Such an ordered arrangement will be called a ranking and the ordinal number indicating the position of a given individual in the ranking is called its rank. A ranking where two or more individuals are allotted the same rank is called a tie. The correlation coefficient computed on the basis of the two series of rank, is called rank-correlation coefficient.

A Spearman's rank correlation coefficient :—

i) The case of no tie :→

C.U.

Suppose we have n individuals ranked according to two characters, A and B.

Individual Serial no.		1	2	3	....	n
Ranks	A	u_1	u_2	u_3	....	u_n
	B	v_1	v_2	v_3	....	v_n

where $(u_1, u_2, \ldots, u_n)$ and $(v_1, v_2, \ldots, v_n)$ are some permutations of $(1, 2, \ldots, n)$. Our problem is to have a suitable measure of the degree of relationship between u and v (or, measure of association between two characters A and B).

Define $d_i = u_i - v_i$, $i = 1(1)n$.

The values of d_i's give an indication of the closeness of the association between A and B.

Spearman suggested as his coefficient of Rank correlation based on $\sum_{i=1}^{n} d_i^2$, the measure

C.V

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}.$$

▨ Rationale :- Smaller the differences $d_i = u_i - v_i,\ i=1(1)n$, higher the association between the characters A and B. Any measure of association should based on d_i's and to ignore the sign of d_i's. We consider $\sum_{i=1}^{n} d_i^2$. Now, $\sum_{i=1}^{n} d_i^2 = 0 \Rightarrow \sum_{i=1}^{n}(u_i-v_i)^2 = 0 \Rightarrow u_i = v_i,\ i=1(1)n$, i.e. there is a perfect positive association or a perfect agreement between the characters if for each individual the ranks in the first and second series would coincide.

Now $\sum_{i=1}^{n} d_i^2$ will be maximum iff the ranking in one character are completely reversed in the other character, i.e. $u_i + v_i = n+1\ \forall i$, $\Rightarrow v_i = n - u_i + 1\ \forall i$, then there is a perfect negative association or a perfect disagreement between the character, and in that case

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n}(u_i - v_i)^2$$

$$= 4\sum_{i=1}^{n}\left(u_i - \frac{n+1}{2}\right)^2 \quad [\because v_i = n - u_i + 1]$$

$$= 4\sum_{i=1}^{n}(u_i - \bar{u})^2 \quad \because \bar{u} = \frac{1}{n}\sum_{i=1}^{n} u_i = \frac{1}{n}\cdot\frac{n(n+1)}{2} = \frac{(n+1)}{2}$$

$$= 4\left\{\sum_{i=1}^{n} u_i^2 - n\bar{u}^2\right\}$$

$$= 4\left\{\frac{n(n+1)(2n+1)}{6} - n\cdot\left(\frac{n+1}{2}\right)^2\right\}$$

$$= \frac{n(n^2-1)}{3}$$

A	B
1	n
2	$n-1$
3	$n-2$
⋮	⋮
n	1

$\boxed{\text{OR}}$

$$\Big] \sum_i d_i^2 = \sum_i (2u_i - n - 1)^2 \qquad [\because v_i = -u_i + n + 1]$$

$$= 4\sum_i u_i^2 - 4(n+1)\sum_i u_i + n(n+1)^2$$

$$= \frac{4\, n(n+1)(2n+1)}{6} - 4(n+1)\cdot\frac{n(n+1)}{2} + n(n+1)^2$$

$$= \frac{n(n^2-1)}{3} \,. \Big]$$

So, when $u_i + v_i = n+1$, we have $\sum_{i=1}^{n} d_i^2 = \frac{n(n^2-1)}{3}$.

Therefore, $0 \le \sum_{i=1}^{n} d_i^2 \le \frac{n(n^2-1)}{3}$

$\Rightarrow 0 \le \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \le 2$, to get a normed measure

$\Rightarrow -1 \le 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \le 1$, to make $\sum_i d_i^2$ and the measure as an increasing function when the characters move from perfect disagreement to perfect agreement. and symmetric about zero.

Hence, r_s is a norming measure and a symmetric measure. It also increases from -1 to 1 when the association goes from perfect disagreement to perfect agreement.

And it is obvious that $r_s = 1$ for the case of perfect agreement and $r_s = -1$ for the case of perfect disagreement between the two series of ranks.

C.U

⇨ Show that → $r_s = -1$ is termed as perfect disagreement.

When there are perfect disagreement then we can consider $u_i + v_i = n+1$, ∀ $i = 1(1)n$.

Here, $\sum_i d_i^2 = \sum_i (u_i - v_i)^2 = \sum_i (2u_i - n - 1)^2$.

$$= 4\sum_i u_i^2 - 4(n+1)\sum_i u_i + n(n+1)^2$$

$$= \frac{4n(n+1)(2n+1)}{6} - 4(n+1)\frac{(n+1)}{2} + n(n+1)^2$$

$$= \frac{n(n^2-1)}{3}.$$

$$\therefore\ r_s = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)} = 1 - \frac{\overset{2}{\cancel{6}}\frac{n(n^2-1)}{\cancel{3}}}{n(n^2-1)} = 1 - 2 = -1.$$

(Proved)

▨ **Result :→** Show that Spearman's Rank correlation coefficient is nothing but the product-moment correlation coefficient between the two series of ranks.

Proof :→ Consider n individuals which are ranked according to two characters A and B.

Serial No.	1	2	3	- - - - - -	n
A	u_1	u_2	u_3	- - - - - -	u_n
B	v_1	v_2	v_3	- - - - - -	v_n

Note that — $\bar{u} = \frac{1}{n}\sum_{i=1}^{n} u_i = \frac{1}{n} \times \frac{n(n+1)}{2} = \frac{n+1}{2}$

similarly, $\bar{v} = \frac{n+1}{2}$.

And, $s_u^2 = \frac{1}{n}\sum_{i=1}^{n}(u_i-\bar{u})^2$

$= \frac{1}{n}\sum_i u_i^2 - \bar{u}^2$

$= \frac{1}{n} \cdot \frac{n(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2$

$= \frac{n^2-1}{12}$.

and, in the same way, $s_v^2 = \frac{n^2-1}{12}$.

Now, we have, $\frac{1}{n}\sum_{i=1}^{n} d_i^2 = \frac{1}{n}\sum_{i=1}^{n}(u_i-v_i)^2$

$= \frac{1}{n}\sum_{i=1}^{n}\{(u_i-\bar{u}) - (v_i-\bar{v})\}^2 \quad [\because \bar{u}=\bar{v}]$

$= \frac{1}{n}\left\{\sum_{i=1}^{n}(u_i-\bar{u})^2 + \sum_i (v_i-\bar{v})^2 - 2\sum_i (u_i-\bar{u})(v_i-\bar{v})\right\}$

$= s_u^2 + s_v^2 - 2\,\mathrm{cov}(u,v)$

so that $\mathrm{cov}(u,v) = \frac{s_u^2 + s_v^2 - \frac{1}{n}\sum_{i=1}^{n} d_i^2}{2}$

$= \frac{n^2-1}{12} - \frac{1}{2n}\sum_i d_i^2$

The correlation coefficient between the two series of ranks is

$$r_{uv} = \frac{\mathrm{cov}(u,v)}{s_u \cdot s_v} = \frac{\frac{n^2-1}{12} - \frac{1}{2n}\sum_i d_i^2}{\frac{n^2-1}{12}} = 1 - \frac{6\cdot\sum_i d_i^2}{n(n^2-1)}.$$

$= r_s$, which is the Spear man's Rank Correlation Coefficient.

Remark :→ By Cauchy-Schwartz inequality,

$$\left\{\sum_i (u_i-\bar{u})(v_i-\bar{v})\right\}^2 \le \left\{\sum_i (u_i-\bar{u})^2\right\}\left\{\sum_i (v_i-\bar{v})^2\right\}$$

$$\Rightarrow \left\{\frac{cov(u,v)}{s_u s_v}\right\}^2 \le 1 \Rightarrow r_{uv}^2 \le 1.$$

$$\Rightarrow r_s^2 \le 1 \Rightarrow \boxed{-1 \le r_s \le 1}.$$

Equality holds iff $r_s = \pm 1$,

iff $u_i - \bar{u} = k(v_i - \bar{v})$ ∀ $i = 1(1)n$.

iff $u_i - \bar{u} = \pm(v_i - \bar{v})$, $i = 1(1)n$. [As, $\sum_i (u_i-\bar{u})^2 = k^2 \sum_i (v_i-\bar{v})^2$
$\Rightarrow s_u^2 = k^2 s_v^2$
$\Rightarrow k^2 = 1$, $\because s_u^2 = s_v^2$]

iff $u_i - v_i = 0$ or $u_i + v_i = n+1$

$\therefore r_s = +1$ iff $u_i - v_i = 0$
iff characters are in perfect agreement.

$\therefore r_s = -1$ iff $u_i + v_i = n+1$
iff characters are in perfect disagreement.

C.V.

ii) The case of ties :→ If the same rank is allotted to k individuals, then we have a tie of length k. If these k individuals follow r other individuals in the ranking, then each may be given the rank r+1.

But we shall follow the convention that each of the k individuals is to be given the rank

$$\frac{(r+1)+(r+2)+\cdots\cdots+(r+k)}{k} = r + \frac{k+1}{2},$$

which is the AM of the ranks that these individuals would have received had there been no ties.

This tie does not affect the mean of the ranks but it affects the variance. The sum of squares of united ranks would be $\sum_{i=1}^{k}(r+i)^2$ but the sum of squares of the tied ranks is

$$k\left\{r + \frac{k+1}{2}\right\}^2.$$

Note that, the difference is

$$\sum_{i=1}^{k}(p+i)^2 - k\left(p+\frac{k+1}{2}\right)^2$$

$$= \sum_{i=1}^{k}\left\{(p+i)-\left(p+\frac{k+1}{2}\right)\right\}^2$$

$$= \sum_{i=1}^{k}\left(i-\frac{k+1}{2}\right)^2$$

$$= \sum_{i=1}^{k} i^2 - k\left(\frac{k+1}{2}\right)^2 \qquad \left[\because \sum_{i=1}^{k} x_i^2 - k\bar{x}^2 = \sum_{i=1}^{k}(x_i-\bar{x})^2\right]$$

$$= \frac{k(k+1)(2k+1)}{6} - k.\frac{(k+1)^2}{4}$$

$$= \frac{k(k^2-1)}{12}$$

Hence, in case of tied ranks, the variance is lowered by $\frac{k^3-k}{12n}$ than the united ranks. Also, it is obvious that the effect of tying different sets is additive.

Now, suppose that there are 's' ties of length $k_1, k_2, \ldots, k_s$ with respect to the first character and there are 't' ties of length $k_1', k_2', \ldots, k_t'$ with respect to the second character. Then the new variances in the case of tied ranks would be

$$s_u'^2 = \frac{n^2-1}{12} - T_u \quad , \text{where } T_u = \sum_{i=1}^{s}\frac{k_i^3-k_i}{12n}$$

and $$s_v'^2 = \frac{n^2-1}{12} - T_v \quad , \text{where } T_v = \sum_{i=1}^{t}\frac{k_i'^3-k_i'}{12n}$$

Similarly, since

$$2\text{cov}(u,v) = s_u'^2 + s_v'^2 - \frac{1}{n}\sum_{i=1}^{n} d_i^2$$

$$\Rightarrow \text{cov}(u,v) = \frac{n^2-1}{12} - \frac{T_u+T_v}{2} - \frac{1}{2n}\sum_{i=1}^{n} d_i^2$$

therefore the spear man's rank correlation coefficient in case of tied ranks reduces to

$$r_s = \frac{\text{cov}(u,v)}{s_u' \cdot s_v'} = \frac{\frac{n^2-1}{12} - \frac{T_u+T_v}{2} - \frac{\sum_{i=1}^{n} d_i^2}{2n}}{\left\{\frac{n^2-1}{12} - T_u\right\}^{1/2}\left\{\frac{n^2-1}{12} - T_v\right\}^{1/2}}$$

▨ Case of $r_s = \pm 1$.

⇒ In case of perfect agreement between the two series of ranks, we shall have $u_i = v_i$, for each i, and hence

$$r_s = \frac{(n^2-1)/12 - T_u}{(n^2-1)/12 - T_u} \qquad \left[\because u_i = v_i \Rightarrow T_u = T_v \text{ and also } \Rightarrow \sum_i d_i^2 = \sum_i (u_i - v_i)^2 = 0 \ \forall\, i \right]$$

$$= 1$$

Again, if there is perfect disagreement between the two sets of ranks, we shall have $v_i = n - u_i + 1$, for each i, and

$$s_u'^2 = s_v'^2 = \frac{(n^2-1)}{12} - T_u .$$

Also, in that case, $T_u = T_v$. Further,

$$\sum_{i=1}^{n} d_i^2/n = \sum_i [2u_i - (n+1)]^2/n = 4s_u^2 = \frac{n^2-1}{3} - 4T_u .$$

As such, we have then

$$r_s = \frac{\frac{n^2-1}{12} - T_u - \left(\frac{n^2-1}{6} - 2T_u\right)}{\frac{n^2-1}{12} - T_u}$$

$$= -1 .$$

▨ Hence, also r_s takes the values -1 and $+1$ in the case of perfect disagreement and perfect agreement between the characters, respectively.

[B] Kendall's rank correlation coefficient :→ [Kendall's τ]

i) The case of no tie :→ Suppose we have

C.U. two series of ranks $(u_1, u_2, \ldots\ldots, u_n)$ and $(v_1, v_2, \ldots\ldots, v_n)$ for n individuals. Consider each possible pair of individuals (i, j), $i<j$ and the rankings of this characters.

Let us define a variable δ_{ij} such that

$$\delta_{ij} = \begin{cases} 1 & \text{if } u_i > u_j \Rightarrow v_i > v_j \\ & \text{and } u_i < u_j \Rightarrow v_i < v_j \\ -1 & \text{if } u_i > u_j \Rightarrow v_i < v_j \\ & \text{and } u_i < u_j \Rightarrow v_i > v_j \end{cases}$$

[In words, if the pair appears in the same order in both the ranking, i.e. if $\{u_i > u_j, v_i > v_j\}$ or $\{u_i < u_j, v_i < v_j\}$, then we assign a score +1 for the pair. But if it appears in the reverse order, i.e. if $\{u_i < u_j, v_i > v_j\}$ or $\{u_i > u_j, v_i < v_j\}$, we assign a score (-1) for the pair.]

C.U. This scores are obtained for each of $\binom{n}{2}$ possible pairs. It is to be noted that → $-\binom{n}{2} \leq \text{total score} \leq \binom{n}{2}$, then we define a rank correlation coefficient τ as

$$\tau = \frac{\text{Total score}}{\text{maximum possible total score}}$$

$$= \frac{\sum_{\substack{i=1 \\ i<j}}^{n} \sum_{j=1}^{n} \delta_{ij}}{\binom{n}{2}}$$, which is known as Kendall's τ.

Features:→

i) ▨ In case of perfect agreement, i.e., if for all the pairs (i,j), $\{u_i > u_j \Rightarrow v_i > v_j\}$ or $\{u_i < u_j \Rightarrow v_i < v_j\}$. The score of each pair is +1, i.e. $\delta_{ij} = +1$ for all the pairs (i,j), and the total score will be $\binom{n}{2}$. Hence, in case of perfect agreement, $\tau = 1$.

On the other hand, if for all the pairs (i,j), $u_i > u_j \Rightarrow v_i < v_j$
or, $u_i < u_j \Rightarrow v_i > v_j$

then $\delta_{ij} = -1$ for all the pairs (i,j) and hence, in case of perfect disagreement $\tau = -1$.

When two characters are independent, $\tau = 0$

ii) ▨ As, $-\binom{n}{2} \leq$ total score $\leq \binom{n}{2}$

$$\Rightarrow -1 \leq \tau = \frac{\text{total score}}{\binom{n}{2}} \leq 1$$

$\Rightarrow$ the coefficient τ is a normed measure and is symmetric.

iii) ▨ τ is a symmetric measure. In case of perfect disagreement, we have $\tau = -1$; when two characters are independent, $\tau = 0$ and in case of perfect agreement, we have $\tau = +1$.

Also $\tau = -0{\cdot}3, +0{\cdot}3$, indicate the same degree of association but in the first case association is negative where as in the second case association is positive. Thus Kendall's τ is symmetric about 0.

iv) ▨ Note that, a pair of individuals $(i,j) : i<j$, can be selected from n, in $\binom{n}{2}$ ways.

Out of $\binom{n}{2}$ pairs, let P be the no. of pairs receiving score '+1' and Q be the no. of pairs receiving score '−1'. Clearly, $P + Q = \binom{n}{2}$.

Thus Kendall's τ reduces to

$$\tau = \frac{P-Q}{\binom{n}{2}}, \text{ since the total score} = (+1)P + (-1)Q = (P-Q).$$

$$= \frac{2P}{\binom{n}{2}} - 1, \text{ or, } \quad 1 - \frac{2Q}{\binom{n}{2}}$$

In practice to calculate the value of τ, we arrange the individuals so that they are in natural order w.r.t. the 1st ranking. In that case, if any pair of individuals are in natural order w.r.t. the 2nd ranking, then for that pair, we take $\delta_{ij} = 1$, otherwise $\delta_{ij} = -1$.

C.U.

▨ Result :→ Kendall's rank correlation coefficient may be regarded as a product moment correlation coefficient.

Proof :→

i) The case of no tie : Consider a pair (i,j) of n individuals, $i<j$. For the first series of ranks $u_1, u_2, \ldots, u_n$, define

$$a_{ij} = \begin{cases} +1 & \text{if } u_i > u_j \\ -1 & \text{if } u_i < u_j \end{cases}$$

Similarly, define,
$$b_{ij} = \begin{cases} +1 & \text{if } v_i > v_j \\ -1 & \text{if } v_i < v_j \end{cases}$$

for the 2nd series of ranks $v_1, v_2, \ldots, v_n$.

Now, $$a_{ij} b_{ij} = \begin{cases} +1 & \text{if } \{u_i > u_j, v_i > v_j\} \text{ or } \{u_i < u_j, v_i < v_j\} \\ -1 & \text{if } \{u_i > u_j, u_i < v_j\} \text{ or } \{u_i < u_j, v_i > v_j\} \end{cases}$$

and
$$\sum_{i<j}\sum a_{ij} b_{ij} = (+1)\{\text{No. of pairs } (i,j),\ i<j \text{ for which } (u_i>u_j,\ v_i>v_j) \text{ or } (u_i<u_j,\ v_i<v_j)\}$$
$$+(-1)\{\text{No. of pairs } (i,j),\ i<j \text{ for which } (u_i>u_j,\ u_i<v_j) \text{ or } (u_i<u_j,\ v_i>v_j)\}$$
$$= (+1)\{\text{the no. of pairs } (i,j),\ i<j \text{ receiving } +1 \text{ score}\}$$
$$+(-1)\{\text{the no. of pairs } (i,j),\ i<j \text{ receiving } -1 \text{ score}\}$$
$$= P - Q$$

$$\sum_{i<j}\sum a_{ij}^2 = \sum_{i<j}\sum b_{ij}^2 = \sum_{i<j} 1 = \binom{n}{2}$$

$$\therefore \tau = \frac{P-Q}{\binom{n}{2}} = \frac{\sum\sum_{i<j} a_{ij} b_{ij}}{\sqrt{\sum\sum_{i<j} a_{ij}^2}\ \sqrt{\sum\sum_{i<j} b_{ij}^2}}$$

which is the product-moment correlation coefficient based on the pairs of values on data $\{(a_{ij}, b_{ij}) : i<j,\ i,j = 1(1)n\}$

C.U

ii) The case of tied ranks:— Let n individuals be ranked w.r.t. two characters. We denote by, $u_1, u_2, \ldots, u_n$, the ranks w.r.t. the 1st character and $v_1, v_2, \ldots, v_n$, the ranks w.r.t. the 2nd character. Let for the (i,j)th pair of individual, where $i<j$, we define —

$$a_{ij} = \begin{cases} +1 & \text{if } u_i > u_j \\ 0 & \text{if } u_i = u_j \\ -1 & \text{if } u_i < u_j \end{cases}$$

for the 1st series of ranks.

Similarly, $$b_{ij} = \begin{cases} +1 & \text{if } v_i > u_j \\ 0 & \text{if } v_i = v_j \\ -1 & \text{if } v_i < v_j \end{cases}$$

for the 2nd series of ranks.

Clearly, $$a_{ij}b_{ij} = \begin{cases} 1 & \text{if } \{u_i > u_j, v_i > v_j\} \text{ or } \{u_i < u_j, v_i < v_j\} \\ 0 & \text{if } u_i = u_j \text{ or } v_i = v_j \\ -1 & \text{if } \{u_i > u_j, v_i < v_j\} \text{ or } \{u_i < u_j, v_i > v_j\} \end{cases}$$

and $$\sum\sum_{i<j} a_{ij}b_{ij} = (+1)\{\text{No. of pairs receiving score } +1\} + (-1)\{\text{No. of pairs receiving score } -1\}$$

$= P - Q$, since we allot score '0' to a pair, if there is a tie on 1st character or on 2nd character or both.

However, in the tied case, if there is a tie of length k, then $\sum\sum_{i<j} a_{ij}^2$ and/or $\sum\sum_{i<j} b_{ij}^2$ are reduced by $\frac{k(k-1)}{2}$.

If there is a tie of length k on the 1st character, the

$$\sum\sum_{i<j} a_{ij}^2 = \binom{n}{2} - \{\text{No. of pairs } (i,j), i<j \text{ receiving '0' score}\}$$
$$= \binom{n}{2} - \binom{k}{2}$$

If there are 's' ties of length $k_1, k_2, \ldots, k_s$ on the 1st character and 't' ties of length $k_1', k_2', \ldots, k_t'$ on the 2nd character.

Then, $$\sum\sum_{i<j} a_{ij}^2 = \binom{n}{2} - \sum_{i=1}^{s}\binom{k_i}{2} = \binom{n}{2} - T_A$$, and

similarly, $$\sum\sum_{i<j} b_{ij}^2 = \binom{n}{2} - \sum_{i=1}^{t}\binom{k_i'}{2} = \binom{n}{2} - T_B$$

So, in case of tied ranks, the formula kendall's τ gives

$$\tau' = \frac{P-Q}{\sqrt{\binom{n}{2} - T_A}\sqrt{\binom{n}{2} - T_B}} = \frac{\sum\sum_{i<j} a_{ij}b_{ij}}{\sqrt{\sum\sum_{i<j} a_{ij}^2}\sqrt{\sum\sum_{i<j} b_{ij}^2}}$$

which is the product moment correlation coefficient.

▨ $\Gamma = \dfrac{\sum\sum_{i<j} a_{ij}b_{ij}}{\sqrt{\sum\sum_{i<j} a_{ij}^2}\sqrt{\sum\sum_{i<j} b_{ij}^2}}$, where a_{ij} and b_{ij} are respectively two scores corresponding to (x_i, x_j) and (y_i, y_j). Then show that $\Gamma = r_{xy}$.

Soln.→ Let us consider, $a_{ij} = x_i - x_j$ and $b_{ij} = y_i - y_j$

$$\therefore \sum\sum_{i<j} a_{ij}^2 = \sum\sum_{i<j} (x_i - x_j)^2 = \frac{1}{2}\sum_i\sum_j (x_i - x_j)^2$$

$$= \frac{1}{2}\left[\sum_i\sum_j \left\{(x_i - \bar{x}) - (x_j - \bar{x})\right\}^2\right]$$

$$= \frac{1}{2}\left[\sum_i\sum_j (x_i - \bar{x})^2 + \sum_i\sum_j (x_j - \bar{x})^2 - 2\sum_i\sum_j (x_i - \bar{x})(x_j - \bar{x})\right]$$

$$= \frac{1}{2}\left[n\sum_i (x_i - \bar{x})^2 + n\sum_j (x_j - \bar{x})^2 - 2\underbrace{\sum_i (x_i - \bar{x})}_{=0}\underbrace{\sum_j (x_j - \bar{x})}_{=0}\right]$$

$$= \frac{1}{2}\left[n.n.s_x^2 + n.n.s_x^2 - 2.0.0\right]$$

$$= \frac{1}{2}2n^2 s_x^2 = n^2 s_x^2$$

Similarly, $\sum\sum_{i<j} b_{ij}^2 = n^2 s_y^2$

Now, $\sum\sum_{i<j} a_{ij}b_{ij} = \sum\sum_{i<j} (x_i - x_j)(y_i - y_j) = \frac{1}{2}\sum_i\sum_j (x_i - x_j)(y_i - y_j)$

$$= \frac{1}{2}\sum_i\sum_j \left[\left\{(x_i - \bar{x}) - (x_j - \bar{x})\right\}\left\{(y_i - \bar{y}) - (y_j - \bar{y})\right\}\right]$$

$$= \frac{1}{2}.2.n^2 \operatorname{cov}(x,y)$$

$$= n^2 \operatorname{cov}(x,y)$$

$$\therefore \Gamma = \frac{n^2 \operatorname{cov}(x,y)}{\sqrt{n^2 s_x^2}\sqrt{n^2 s_y^2}} = \frac{\operatorname{cov}(x,y)}{s_x s_y} = r_{xy}$$

INTRACLASS CORRELATION :

Sometimes specially in biological situation, it is desired to study the correlation between some characteristics of members of one or more families. thus we may be interested in the correlation of heights of brothers or weights of sisters in a family. By correlation, here we mean the extent to which the members of the same family or group resemble (are related) to each other w.r.t. a given characteristic or variable. Such a correlation, we shall call Intra-class Correlation. (c.v.)

Let us consider that there are p families (groups) each consisting of k individuals. Suppose x_{ij} is the value for the j^{th} member of the i^{th} family (group) on the character (or, variable), where, $j=1(1)k$, $i=1(1)p$.

Now consider a pair (j,j'), $j\neq j'$ of members from the ith family, $i=1(1)p$, $j\neq j'=1(1)k_i$, then we have a bivariate data

$$\{(x_{ij}, x_{ij'}) : i=1(1)p,\ j\neq j'=1(1)k_i\}$$, cosisting of

$$N=\sum_{i=1}^{p} k_i(k_i-1)$$ pairs of values.

The product-moment correlation coefficient computed from the above bivariate data is called the Intra-class correlation coefficient.

Let the pair of values $(x_{ij}, x_{ij'})$ be an observed value on the pair of variables, say, (u,v). Hence the intra-class correlation coefficient is given by

$$r_I = r_{uv} = \frac{cov(u,v)}{s_u \cdot s_v}.$$

These are the concepts of Intra-class correlation.

In general case, when there are k_i members in the ith class, in the correlation table each member of the ith class will appear $(k_i - 1)$ times in each column in association with the other members of the class. Here —

$$\overline{U} = \overline{V} = \frac{\sum_{i=1}^{p}\sum_{j=1}^{k_i}(k_i-1)x_{ij}}{\sum_{i=1}^{p} k_i(k_i-1)}$$

$$= \frac{1}{N}\sum_{i=1}^{p}\left\{(k_i-1)\sum_{j=1}^{k_i} x_{ij}\right\} = \overline{x}_0 \text{, say}$$

(which is not grand mean of x)

where, $N = \sum_{i=1}^{p} k_i(k_i-1)$.

Similarly, $s_u^2 = s_v^2 = \frac{1}{N}\sum_{i=1}^{p}\left\{(k_i-1)\sum_{j=1}^{k_i}(x_{ij}-\overline{x}_0)^2\right\}$.

Again, $\mathrm{Cov}(u,v)$ is equal to,

$$\mathrm{Cov}(u,v) = \frac{1}{N}\sum_{i=1}^{p}\sum_{j,j'=1}^{k_i}(x_{ij}-\overline{x}_0)(x_{ij'}-\overline{x}_0) \text{, with } j \neq j'.$$

$$= \frac{1}{N}\sum_{i=1}^{p}\sum_{j,j'=1}^{k_i}{}'(x_{ij}-\overline{x}_0)(x_{ij'}-\overline{x}_0) - \frac{1}{N}\sum_{i=1}^{p}\sum_{j=1}^{k_i}(x_{ij}-\overline{x}_0)^2$$

(where $\sum'$ extends over all possible pairs, including the case $j = j'$)

$$= \frac{1}{N}\sum_i k_i^2(\overline{x}_i-\overline{x}_0)^2 - \frac{1}{N}\sum_i\sum_j(x_{ij}-\overline{x}_0)^2.$$

where $\overline{x}_i$ is the mean of the ith class, being equal to $\frac{1}{k_i}\sum_{j=1}^{k_i} x_{ij}$.

Hence, $$r_I = \frac{\mathrm{cov}(U,V)}{\sqrt{\mathrm{Var}(U)\mathrm{Var}(V)}} = \frac{\sum_i k_i^2(\overline{x}_i-\overline{x}_0)^2 - \sum_i\sum_j(x_{ij}-\overline{x}_0)^2}{\sum_i\left\{(k_i-1)\sum_j(x_{ij}-\overline{x}_0)^2\right\}}$$

This is the intra-class correlation coefficient for the general case.

(c.v.)

Case of eaual no. of members in each family :

If there are K indivisuals (members) in a group (family), there will be K(K-1) pairs. thus if we have p groups (families) of k individuals (members) in each, there will be pk(K-1) pairs of values in the correlation table. Let x_{ij} denote the variate value for the jth member of the i-th family $[i=1,2,\ldots,p;\ j=1,2,\ldots,k]$.

If we arbitarily regard the first column of the correlation table as corresponding to a variate U and the second corresponding to a variate V, then the mean of each variate is given by,

$$\bar{U}=\bar{V}=\frac{1}{pk(k-1)}\sum_{i=1}^{p}(k-1)\sum_{j=1}^{k}x_{ij}$$

$$=\frac{1}{pk}\sum_i\sum_j x_{ij}=\bar{x}\text{, the grand mean of } x.$$

Since each of the values x_{ij} occurs $(k-1)$ times in each column of the correlation table, along with the values for the others $(k-1)$ members of the family occuring in the others column.

Similarly, the variance of each variate is given by

$$s_U^2=s_V^2=\frac{1}{pk(k-1)}\sum_{i=1}^{p}\quad(k-1)\sum_{j=1}^{k}(x_{ij}-\bar{x})^2$$

$$=\frac{1}{pk}\sum_i\sum_j(x_{ij}-\bar{x})^2$$

$$=s^2\text{, the total variance of } x.$$

Now,
$$\text{cov}(U,V)=\frac{1}{pk(k-1)}\sum_{i=1}^{p}\sum_{\substack{j,j'=1\\ j\neq j'}}^{k}(x_{ij}-\bar{x})(x_{ij'}-\bar{x})$$

$$=\frac{k^2}{pk(k-1)}\sum_{i=1}^{p}(\bar{x}_i-\bar{x})^2-\frac{1}{pk(k-1)}\sum_{i=1}^{p}\sum_{j=1}^{k}(x_{ij}-\bar{x})^2$$

writting $s_m^2=\frac{1}{p}\sum_{i=1}^{p}(\bar{x}_i-\bar{x})^2$, the variance of the means of the p families, then we have → $\text{cov}(U,V)=\frac{k}{(k-1)}s_m^2-\frac{s^2}{k-1}$.

Now, the coefficient of intra-class correlation, r_I, is given by

$$r_I=\frac{\text{cov}(U,V)}{\sqrt{\text{var}(U)\,\text{var}(V)}}$$

$$=\frac{\frac{k}{k-1}s_m^2-\frac{s^2}{k-1}}{s^2}$$

$$=\frac{1}{k-1}\left\{k.\frac{s_m^2}{s^2}-1\right\}$$

c.v. **Properties :→**

i) r_I lies between $-\frac{1}{k-1}$ to 1.

Proof :→

$$s^2 = \frac{1}{pk}\sum_{i=1}^{p}\sum_{j=1}^{k}(x_{ij}-\bar{x})^2$$

$$= \frac{1}{pk}\sum_{i=1}^{p}\sum_{j=1}^{k}(x_{ij}-\bar{x}_i+\bar{x}_i-\bar{x})^2$$

$$= \frac{1}{pk}\sum_{i=1}^{p}\sum_{j=1}^{k}(x_{ij}-\bar{x}_i)^2+\frac{1}{pk}\sum_{i=1}^{p}\cancel{k}(\bar{x}_i-\bar{x})^2$$

[∵ The product term $= \frac{2}{pk}\sum_{i=1}^{p}(\bar{x}_i-\bar{x})\left\{\sum_{j=1}^{k}(x_{ij}-\bar{x}_i)\right\}$

$= 0$]

$$= \frac{1}{pk}\sum_{i=1}^{p}\sum_{j=1}^{k}(x_{ij}-\bar{x}_i)^2+\frac{1}{p}\sum_{i=1}^{p}(\bar{x}_i-\bar{x})^2$$

$$\therefore s^2 = \frac{1}{pk}\sum_{i=1}^{p}\sum_{j=1}^{k}(x_{ij}-\bar{x}_i)^2+s_m^2 = s_w^2+s_m^2$$

Clearly, $0 \le s_m^2 \le s^2$

$$\Rightarrow 0 \le \frac{s_m^2}{s^2} \le 1$$

$$\Rightarrow 0-1 \le \frac{k s_m^2}{s^2}-1 \le k-1$$

$$\Rightarrow -\frac{1}{k-1} \le \frac{1}{k-1}\left\{\frac{k s_m^2}{s^2}-1\right\} \le 1$$

$$\Rightarrow -\frac{1}{k-1} \le r_I \le 1.$$

i.e. r_I lies between $-\frac{1}{k-1}$ to 1.

[Note :- $s_m^2 = \frac{1}{p}\sum_{i=1}^{p}(\bar{x}_i-\bar{x})^2$ is called the variance between the p-families. and,

$s_w^2 = \frac{1}{pk}\sum_{i=1}^{p}\sum_{j=1}^{k}(x_{ij}-\bar{x}_i)^2$ is called the variance within the families.]

(ii) (a) r_I is maximum, i.e. $r_I = +1$.

iff $s_w^2 = \frac{1}{pk}\sum_{i=1}^{p}\sum_{j=1}^{k}(x_{ij}-\bar{x}_i)^2 = 0$, i.e. variance within families is zero.

iff $\{x_{ij} = \bar{x}_i \;\forall\; j=1(1)k,\ \forall\; i=1(1)p\}$, i.e. the variate values within each family are all equal.

(b) r_I is minimum, i.e. $r_I = -\frac{1}{k-1}$.

iff $s_m^2 = \frac{1}{p}\sum_{i=1}^{p}(\bar{x}_i-\bar{x})^2 = 0$, i.e. the variance between the families is zero.

iff $\bar{x}_i = \bar{x} \;\forall\; i=1(1)p$. , i.e. there is no variability between family members.

C.U.

(iii) r_I is not a symmetric measure, or, r_I is a skewed measu

As $-\frac{1}{k-1} \le r_I \le 1$, the lower limit $-\frac{1}{k-1}$ is larger than -1 unless $k=2$. The lower limit also varies with varying k. Hence, caution is necessary in the interpretation of the intra-class correlation coefficient.

For example, consider the case; $k=5$, then $-\frac{1}{4} \le r_I \le 1$. Here, $r_I = -\frac{1}{4}$ means that there is a minimum intra-class correlation, whereas $r_I = +\frac{1}{4}$ means that there is a positive low intra-class correlation, since the maximum value of r_I is $+1$. It is thus skew-coefficient; i.e. not a symmetric coefficient in the sense that a negative value of it has not the same significance in terms of extent of association as an equivalent positive value.

In fact, if k is very large, r_I is almost always positive.

Distinguish between Intra-class & Inter-class Correlation Coefficient

(1) The ordinary correlation coefficient between two variables, e.g. the correlation between the height of tallest and shortest brother could be called an inter-class correlation coefficient. By intra-class correlation coefficient, we mean the extent to which the members of a family or group resemble to each other.

(2) Let there be p families containing $k_1, k_2, \ldots, k_p$ members and x_{ij}, the value of the j^{th} member of the ith family. The inter-class correlation coefficient computed from the data $\{(x_{it}, x_{is}) : i = 1(1)p\}$ where x_{it}, x_{is} denote the heights of the tallest and (smallest) shortest in the i^{th} family, is the simple correlation coefficient

$$r_{ts} = \frac{\sum_i (x_{it} - \bar{x}_t)(x_{is} - \bar{x}_s)}{\sqrt{\sum_i (x_{it} - \bar{x}_t)^2} \sqrt{\sum_i (x_{is} - \bar{x}_s)^2}}$$

The intra-class correlation coefficient is obtained from the data $\{(x_{ij}, x_{ij'}) : j \neq j' = 1(1)k_i, i = 1(1)p\}$ cosisting of $N = \sum_{i=1}^{p} k_i(k_i - 1)$ values as —

$$r_I = \frac{\sum_{i=1}^{p} k_i^2 (\bar{x}_i - \bar{x}_0)^2 - \sum_{i=1}^{p} \sum_{j=1}^{k_i} (x_{ij} - \bar{x}_0)^2}{\sum_{i=1}^{p} \left\{ (k_i - 1) \sum_{j=1}^{k_i} (x_{ij} - \bar{x}_0)^2 \right\}}$$

(3) We have, $-1 \leq r_{ts} \leq 1$ and r_{ts} is a symmetric measure. We have, $-\frac{1}{k-1} \leq r_I \leq 1$ and r_I is not a symmetric measure, i.e. r_I is a skewed measure.

— x —

CHAPTER-7

Categorical Data Analysis

Categorical Variable and Categorical Data : ~ A variable which takes values on a scale consisting of some categories is called a Categorical variable. For example, "Political Idealogy" is a categorical variable which takes values on a scale consisting of several categories, say, 'liberal', 'moderate' & 'conservative'. Any data collected on a categorical variable is called Categorical Data.

Response & Explanatory Variables: Most statistical techniques distinguish between response or dependent variables and explanatory or independent variables. For instance, regression models describe how the mean of response variable such as the selling prices of houses changes according as the values of explanatory variable such as square footage and location.

Nominal & Ordinal Variables : Categorical variables have two primary types of scales. Variables having categories without a natural ordering are called nominal variable. e.g. religious affiliation with the categories catholic, protestant, jewish, Muslims and others. For nominal variables, the order of listing of the categories is irrelevant (not necessary). Many categorical variables do have ordered categories, such variables are called ordinal variables. Examples are social class: upper, middle, lower; & patient conditions: good, fair, serious.

Attribute : ~ When we record the sex of each newborn baby during a month or the languge of each book in a library, the data are not numbers initially. We get numbers if subseauently, we note the number of male babies and that of female babies, or the numbers of books written in English, the number written in Hindi, the numbers written in Bengali and so forth. For this type of data, the characters observed is not expressible in numerical terms. Such a character is, therefore, called a qualitative characters or an attribute.

Dichotomy :- A classification of the simple kind considered, in which each class is divided into two sub-classes and no more, has been termed by Logicians classification, or to use the more strictly applicable term, division by dichotomy. The classification of most statistics are not dichotomous, for most usually a class is divided into more than two subclasses, but dichotomy is the fundamental case.

Notations :- For theoretical purposes, it is necessary to have some simple notations for the classes formed and for the numbers of observation assigned to each. Let us denote the several attributes by A, B, C, ..., etc. An object or an individual possessing the attribute A will be termed simply by A. The class, all the members of which possess the attribute A will be termed as the class-A. The absence of the attributes A, B, C, ...; we shall employ the notations $\alpha, \beta, \gamma, \ldots$ accordingly. Thus, if A represents the attribute blindness; α represents sight. If B stands for deafness, β stands for hearing.

Generally, 'α' is equivalent to 'not A'. The class-α is equivalent to the class none of the members of which possesses the attribute A.

In case of combination of attributes, such as, 'AB' represents the combination of blindness and deafness. If a third attribute be noted, e.g. – insanity, denoted by C & the absence of it by γ; then the class 'ABC' includes those who are at once deaf, blind and insane; 'ABγ' includes those who are deaf and blind but not insane.

Class-frequencies :- The no. of observation assigned to any class is termed as the 'class-frequency'. Class-frequencies will be denoted by putting the class-symbols within parenthesis.

Orders of classes and class-frequency :-

For 2 attributes, the class frequencies of different orders are

Order 0 : N

1st order : (A), (B),
(α), (β)

[(A) denotes the class frequency of the class A]

2nd order : (AB), (Aβ)
(αB), ($\alpha\beta$)

Any class frequency can always be expressed in terms of frequencies of higher order classes. e.g.

$$(A) = (AB) + (A\beta)$$
$$(AB) = (ABC) + (AB\gamma)$$

Ultimate Class frequencies :- A class of heighest order is known as an ultimate class and its frequency an ultimate class frequency.

Every class frequency can be written as a sum of certain ultimate class frequencies.

For any frequency can be analysed into higher frequencies and the process needs stop only when we have reached the frequency of the highest order, e.g. with 3 attributes,

$$(A) = (AB) + (A\beta)$$
$$= (ABC) + (AB\gamma) + (A\beta C) + (A\beta\gamma)$$

The total no. of class-frequencies :—

Order zero :	N		
Order one :	(A)	(B)	(C)
	(α)	(β)	(γ)
Order two :	(AB)	(AC)	(BC)
	(Aβ)	(Aγ)	(Bγ)
	(αB)	(αC)	(βC)
	(αβ)	(αγ)	(βγ)
Order three :	(ABC)	(αBC)	
	(ABγ)	(αBγ)	
	(AβC)	(αβC)	
	(Aβγ)	(αβγ)	

Thus, here we have 3^3 distinct class-frequencies in all. In general, for n attributes there are 3^n distinct class-frequencies, provided we count N as a frequency of order zero.

Of order zero, there is a single class N; of order one, there are 2n classes; of order two, there are $\binom{n}{2}2^2$ classes; hence, of order r, there are $\binom{n}{r}2^r$ classes.

∴ No of class frequencies $= \sum_{r=0}^{n}\binom{n}{r}2^r = (1+2)^n = 3^n$

▨ **Consistency :→** Any class frequencies which have been or might have been observed within one and the same population may be said to be consistent with one another. They conform with one another, and don't, in any way, conflict.

Symbols :> We define, $A.N = (A)$, for any attribute

similarly, $\alpha.N = (\alpha)$

Now, $(A+\alpha)N = (A)+(\alpha) = N$

$$\Rightarrow (A+\alpha) = 1$$
$$\Rightarrow \alpha = 1-A$$

Therefore, in any symbolic expression we can replace the operators A by $1-\alpha$ or α by $1-A$.

Again, $(AB) = A.(B) = B.(A)$

A little reflection will show that the operative symbols therefore obey the laws of algebra.

$$\therefore (\alpha\beta) = \alpha\beta.N$$
$$= (1-A).(1-B).N$$
$$= (1-B-A+AB).N$$
$$= N-(A)-(B)+(AB)$$

Similarly, $(\alpha\beta\gamma) = \alpha\beta\gamma.N$
$$= (1-A)(1-B)(1-C).N$$
$$= (1-A-B-C+AB+BC+AC-ABC).N$$
$$= N-(A)-(B)-(C)+(AB)+(BC)+(AC)-(ABC)$$

Condition for Consistency :— The attributes denoted by capitals may be termed positive attributes, on the other hand, attributes denoted by Greek letters, may be termed as negative attributes.

The necessary and sufficient condition for the consistency of a set of independent class-frequencies is that no ultimate class-frequency be negative.

If two attributes are noted, there are 4 ultimate frequencies – (AB), $(A\beta)$, (αB), $(\alpha\beta)$. Expressing them in terms of positive classes, we find the following conditions —

i) $(AB) \geq 0$

ii) $(AB) \geq (A)+(B)-N$ [i.e. $(\alpha\beta) \geq 0$]

iii) $(AB) \leq (A) \Leftrightarrow (A\beta) \geq 0$

iv) $(AB) \leq (B) \Leftrightarrow (\alpha B) \geq 0$

For 3 attributes the conditions that the 8 ultimate frequencies are not negative will be found to lead to the following —

i> $(ABC) \geq 0$

ii> $(ABC) \geq (AB)+(AC)-(A)$

iii> $(ABC) \geq (AB)+(AC)-(B)$

iv> $(ABC) \geq (AB)+(AC)-(C)$

v> $(ABC) \leq (AB)$

vi> $(ABC) \leq (AC)$

vii> $(ABC) \leq (BC)$

viii> $(ABC) \leq N-(A)-(B)-(C)+(AB)+(BC)+(AC)$

Example:> Show that neither x nor y can exceed 1/4 when the followings are given:

$$\frac{(A)}{N}=x,\ \frac{(B)}{N}=2x,\ \frac{(C)}{N}=3x \ \&\ \frac{(AB)}{N}=\frac{(BC)}{N}=\frac{(CA)}{N}=y.$$

Soln.> From condition of consistency,

$$(AB) \leq (A)$$

$$\Rightarrow y \leq x \quad \text{———} (*)$$

$$(BC) \geq (B)+(C)-N \Rightarrow \frac{(BC)}{N} \geq \frac{(B)}{N}+\frac{(C)}{N}-1$$

$$\Rightarrow y \geq 2x+3x-1$$

$$\Rightarrow y \geq 5x-1 \quad [\text{Using } (*)]$$

$$\Rightarrow x \geq 5x-1$$

$$\Rightarrow x \leq 1/4$$

$$\Rightarrow y \leq x \leq 1/4.$$

INDEPENDENCE:

Attribute	B	β	TOTAL
A	(AB)	(Aβ)	(A)
α	(αB)	(αβ)	(α)
TOTAL	(B)	(β)	N

Two related attributes A & B are said to be independent if

$$\frac{(AB)}{(B)}=\frac{(A\beta)}{(\beta)}$$

$$\Leftrightarrow \frac{(AB)}{(B)}=\frac{(AB)+(A\beta)}{(B)+(\beta)}=\frac{(A)}{N}$$

$$\Leftrightarrow \frac{(AB)}{N}=\frac{(A)}{N}\cdot\frac{(B)}{N} \quad \text{———} (**)$$

"If the attributes A and B are independent, the proportion of AB's in the population is equal to the proportion of A's multiplied by the proportion of B's." → fundamental rule.

If there is no sort of relationship of any kind between two attributes A and B, we expect to find the same proportion of A's amongst the B's as amongst the not B's.

We have earlier the criterion of independence for A & B –

$$\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)} \qquad \text{———} (*)$$

If this relation holds good, then the following equations must also hold —

$$\frac{(\alpha B)}{(B)} = \frac{(\alpha\beta)}{(\beta)}$$

$$\frac{(AB)}{(A)} = \frac{(\alpha B)}{(\alpha)}$$

$$\frac{(A\beta)}{(A)} = \frac{(\alpha B)}{(\alpha)}$$

$$(*) \longrightarrow \text{gives} \rightarrow \quad \frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)}$$

$$\Leftrightarrow \frac{(B)-(AB)}{(B)} = \frac{(\beta)-(A\beta)}{(\beta)}$$

$$\text{i.e. } \frac{(\alpha B)}{(B)} = \frac{(\alpha\beta)}{(\beta)}, \text{ etc.}$$

NOTE: If 2 attributes A & B are independent, then i) α, β are also independent, ii) A, β are also independent.

i) For,
$$\begin{aligned}(\alpha\beta) &= N-(A)-(B)+(AB)\\ &= N-(A)-(B)+\frac{(A)(B)}{N}\\ &= N-(A)-\frac{(B)}{N}\left(N-(A)\right)\\ &= \frac{(N-(A))(N-(B))}{N}\\ &= \frac{(\alpha)(\beta)}{N}\end{aligned}$$

ii)
$$\frac{(A\beta)}{(\beta)} = \frac{(AB)+(A\beta)}{(B)+(\beta)} = \frac{(A)}{N}$$

$$\Rightarrow (A\beta) = \frac{(A)(\beta)}{N}$$

Interpretation: → The advantages of (**) over (*) is that it gives expressions for the second order frequency in terms of the frequencies of the 1st order and the total no. of observations alone.

The criteria of independence may be expressed in yet a 3rd form, i.e. in terms of 2nd order frequencies alone.

If A and B are independent then

$$(AB)\cdot(\alpha\beta) = \frac{(A)(B)}{N}\cdot\frac{(\alpha)(\beta)}{N} = \frac{(A)(B)(\alpha)(\beta)}{N^2}$$

$$\Rightarrow (AB)\cdot(\alpha\beta) = (\alpha B)(A\beta)$$

$$\Leftrightarrow \frac{(AB)}{(\alpha B)} = \frac{(A\beta)}{(\alpha\beta)} \text{ ——— ①}$$

$$\text{or, } \frac{(AB)}{(A\beta)} = \frac{(\alpha B)}{(\alpha\beta)} \text{ ——— ②}$$

① may be read : " The ratio of A's to α's amongst the B's is equal to the ratio of A's to α's amongst the β's.

② can be intepreted in a similar way.

Example :-> If the second-order frequencies have the following values, are A and B independent or not ?

$(AB) = 110$, $(\alpha B) = 90$, $(A\beta) = 290$, $(\alpha\beta) = 510$

Soln. -> Clearly, $(AB)(\alpha\beta) = 110 \times 510$

$(A\beta)(\alpha B) = 290 \times 90$

so, $(AB)(\alpha\beta) > (\alpha B)(A\beta)$

So, A and B are not independent.

ASSOCIATION :-> A and B are said to be positively associated (or, simply associated) if

$$(AB) > \frac{(A)\cdot(B)}{N}$$

A and B are said to be negatively associated (or, simply disassociated) if

$$(AB) < \frac{(A)\cdot(B)}{N}$$

In common language, one speaks of A and B as being 'associated' if they appear together in a number of cases. But here, by positive assciation between A and B, we mean that they appear together in a greater no. of cases than is to be expected if they are independent.

Complete Association or Disassociation :-

The circumstances in which the association between two attributes is said to be complete, can be explained into two cases ——

i) We may say that for complete association, all A's must be B's and all B's must be A's, i.e. $[(A\beta)=0]$ & $[(\alpha B)=0]$ or otherwise. Therefore, we see that A's and B's occur in the population in equal numbers.

ii) We may adopt a rather wider meaning and say that all A's are B's or all B's are A's, according to whether the A's or the B's are in the minority.

Similarly, complete disassociation may be taken either as the case when no A's are B's and no α's are β's, or more widely as the case when either of these statements is true.

Thus two attributes are completely associated if one of them can't occur without the other, though the other may occur without the one.

The symbols $(AB)_0$ and δ := We define $(AB)_0 = \frac{(A)(B)}{N}$, the value of (AB) under the assumption that the attributes are independent.

We shall use the other symbols
$$(\alpha\beta)_0 = \frac{(\alpha)(\beta)}{N},\quad (\alpha B)_0 = \frac{(\alpha)(B)}{N},\quad (A\beta)_0 = \frac{(A)(\beta)}{N}$$

If δ denote the excess of (AB) over $(AB)_0$, then keeping the marginal totals fixed, the table reduces to the form—

Attribute	B	β	TOTAL
A	$(AB)_0+\delta$	$(A\beta)_0-\delta$	(A)
α	$(\alpha B)_0-\delta$	$(\alpha\beta)+\delta$	(α)
TOTAL	(B)	(β)	N

Define, $\delta = (AB)-(AB)_0$ and also, quite generally we have—

$$\delta = (AB)-(AB)_0 = (\alpha\beta)-(\alpha\beta)_0 = (A\beta)_0-(A\beta) = (\alpha B)_0-(\alpha B)$$

Note that, $\delta = (AB) - (AB)_0 = (AB) - \frac{(A)(B)}{N}$

$$\delta = \frac{N(AB) - (A)(B)}{N}$$

$$\Rightarrow \delta = \frac{1}{N}\left[N(AB) - \{(AB) + (A\beta)\}\{(AB) + (\alpha B)\}\right]$$

$$= \frac{1}{N}\left[\{(AB) + (A\beta) + (\alpha B) + (\alpha\beta)\}(AB) - \{(AB) + (A\beta)\}\{(AB) + (\alpha$$

$$= \frac{1}{N}\left[(AB)(\alpha\beta) - (A\beta)(\alpha B)\right]$$

Clearly, $\delta = 0$, $\delta > 0 \Leftrightarrow$ (+ve) assciation, $\delta < 0 \Leftrightarrow$ (−ve) association

$\Leftrightarrow$ A and B are independent.

'δ' determines uniquely the departure from independence. But it may be of interest to measure the intensity of association. It's called 'δ measure' for association between 2 attributes

▨ Coefficient of association :–

$$\delta = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{N}$$

1) Yule's coefficient of association :– As a measure of intensity of association between 2 attributes A and B, Yule gave the following coefficient of association:

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

$$= \frac{N\delta}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

If A and B are independent, $\delta = 0 \Leftrightarrow Q = 0$.

*

2>	A	α
B	k(AB)	αB
β	k(Aβ)	$\alpha\beta$
	k(A)	(α)

PROPERTIES :–

1) $Q = \frac{N\delta}{(AB)(\alpha\beta) + (\alpha B + A\beta)}$ increases as δ increases.

* 2) If all the term containing A are multiplied by a constant, the value of Q remains unaltered. Similar things hold for α, β and B.

Hence, Q is independent of A's and α's in the data.

∴ Relative proportion of A's and α's are $\frac{k(A)}{k(A)+\alpha}$, $\frac{(\alpha)}{k(A)+(\alpha)}$

Now, $Q^* = \frac{k(AB)(\alpha\beta) - k(A\beta)(\alpha B)}{k(AB)(\alpha\beta) + k(A\beta)(\alpha B)} = Q$

Hence, Q is independent of the relative proportion of A's.

N.P. Q is a symmetric measure (coefficient) and it increases from −1 to 1 as the extent of association increases from perfect −ve to perfect +ve. Shown in property 3).

3> $-1 \le Q \le 1$

Proof:→ $Q = \dfrac{(AB)(\alpha\beta) - (\alpha B)(A\beta)}{(AB)(\alpha\beta) + (\alpha B)(A\beta)} = \dfrac{a-b}{a+b}$,

where $a = (AB)(\alpha\beta) \ge 0$ & $b = (\alpha B)(A\beta) \ge 0$

Note that, $Q = 1 - \dfrac{2b}{a+b} \le 1$, —①

and, $Q = -1 + \dfrac{2a}{a+b} \ge -1$ —②

Combining ① & ② ⇒ $-1 \le Q \le 1$.

Marginal cases :→ $Q = 0 \Leftrightarrow \delta = 0$ iff A & B are independent.

$Q = 1$ iff $b = 0$

$\Leftrightarrow (\alpha B)(A\beta) = 0$

$\Leftrightarrow (\alpha B) = 0$, or, $(A\beta) = 0$

$\Leftrightarrow (AB) = (B)$, or, $(AB) = (A)$

⇔ A & B are in complete association,

$Q = -1$ iff A & B are in complete disassociation. O.U.

C.U.

2> Yule's coefficient of colligation :

Define, $Y = \dfrac{1 - \sqrt{(A\beta)(\alpha B)/(AB)(\alpha\beta)}}{1 + \sqrt{(A\beta)(\alpha B)/(AB)(\alpha\beta)}}$

$= \dfrac{\sqrt{(AB)(\alpha\beta)} - \sqrt{(A\beta)(\alpha B)}}{\sqrt{(AB)(\alpha\beta)} + \sqrt{(A\beta)(\alpha B)}}$

▨ Show that → $Q = \dfrac{2Y}{1+Y^2}$

Proof:→ $Y = \dfrac{\sqrt{(AB)(\alpha\beta)} - \sqrt{(A\beta)(\alpha B)}}{\sqrt{(AB)(\alpha\beta)} + \sqrt{(A\beta)(\alpha B)}}$

$\Rightarrow \dfrac{1+Y}{1-Y} = \dfrac{\sqrt{(AB)(\alpha\beta)}}{\sqrt{(A\beta)(\alpha B)}}$

$\Rightarrow \dfrac{(1+Y)^2}{(1-Y)^2} = \dfrac{(AB)(\alpha\beta)}{(A\beta)(\alpha B)}$

$\Rightarrow \dfrac{(1+Y)^2 - (1-Y)^2}{(1+Y)^2 + (1-Y)^2} = \dfrac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$

$\Rightarrow \dfrac{4Y}{2(1+Y^2)} = Q$

$\Rightarrow Q = \dfrac{2Y}{1+Y^2}$.

N.P. Y has the same properties as Q, i.e. $-1 \le Y \le 1$. '=' holds, i.e. (+1) for complete association & (-1) for complete disassociation. If A & B are independent, then Y vanishes.

⇨ **Manifold Classification** :→ In stead of dividing the population under consideration into two parts by a simple dechotomy, we may also divide it into a number of parts by a simple process. For instance, we can extend the dichotomy of the population of men into " those with blue eyes" and " those not with blue eyes" to a threefold division: " those with blue eyes", "those with brown eyes" and " those with neither blue nor brown eyes"; or into a fourfold division by adding a fresh category, " those with grey eyes";

Generally, our population may be divided first according to s heads, $A_1, A_2, \ldots\ldots, A_s$; each of the classes so obtained into t heads, $B_1, B_2, \ldots\ldots, B_t$; each of these into u heads, $C_1, C_2, \ldots\ldots, C_u$; and so on.

This is called "manifold classification."

The theory of manifold classification involving n attributes is rather complicated, but its fundamental principles are very similar to dichotomy; A straight forward extension of the methods already discussed will give the following results:

(a) There are $s \times t \times u \times \ldots\ldots$ ultimate classes.

(b) The total no. of classes, including N and the ultimate classes, is $(s+1)(t+1)(u+1)\ldots\ldots$

(c) The data are consistent if, and only if, every ultimate class-frequency is not negative.

(d) The data are completely specified by $s \times t \times u \times \ldots\ldots$ algebrically independent class frequencies. Even if all these are not given, it may be possible to set limits to the other class-frequencies.

▨ **Contingency Table** :— Let us consider a classification in which the attribute A is s-fold and B is t-fold. Clearly, here we have 'st' classes of the type A_iB_j ; $i=1(1)s$, $j=1(1)t$.

Then we can arrange our data in a table as follows —

B \ A	A_1	A_2	$\cdots$	A_m	$\cdots$	A_s	Totals
B_1	(A_1B_1)	(A_2B_1)	$\cdots$	(A_mB_1)	$\cdots$	(A_sB_1)	(B_1)
B_2	(A_1B_2)	(A_2B_2)	$\cdots$	(A_mB_2)	$\cdots$	(A_sB_2)	(B_2)
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
B_n	(A_1B_n)	(A_2B_n)	$\cdots$	(A_mB_n)	$\cdots$	(A_sB_n)	(B_n)
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
B_t	(A_1B_t)	(A_2B_t)	$\cdots$	(A_mB_t)	$\cdots$	(A_sB_t)	(B_t)
Totals	(A_1)	(A_2)	$\cdots$	(A_m)	$\cdots$	(A_s)	N

Such a table is called a Contingency Table.

Alternative Notations:

$$n_{ij} = (A_iB_j)$$
$$n_{io} = (A_i)$$
$$n_{oj} = (B_j)$$

Coefficients of contingency: The characteristics A & B are said to be completely independent in the population at large, if for all values of m and n, we must have —

$$(A_mB_n) = \frac{(A_m)(B_n)}{N} = (A_mB_n)_0 \quad \text{, i.e., } n_{ij} = \frac{n_{io}\cdot n_{oj}}{N}$$

If, however, A and B are not completely independent, (A_mB_n) and $(A_mB_n)_0$ will not be identical for all the values of m and n. The deviation from independence in that particular cell will be measured by,

$$\delta_{mn} = (A_mB_n) - (A_mB_n)_0$$
$$= (A_mB_n) - \frac{(A_m)(B_n)}{N}$$
$$\text{i.e., } \delta_{ij} = n_{ij} - \frac{n_{io}\cdot n_{oj}}{N}$$

Properties :→

(1) In General $\delta_{mn} \neq \delta_{nm}$.

(2) the δ's are not all algebraically independent:

We have, in fact, for any particular m,

$$\delta_{m1} + \delta_{m2} + \dots\dots + \delta_{mn} + \dots\dots + \delta_{mt}$$

$$= \sum_{j=1}^{t} \delta_{mj}$$

$$= (A_m B_1) - \frac{(A_m)(B_1)}{N} + (A_m B_2) - \frac{(A_m)(B_2)}{N} + \dots\dots + (A_m B_t) - \frac{(A_m)(B_t)}{N}$$

$$= (A_m) - \frac{(A_m)}{N}\left\{(B_1) + (B_2) + \dots\dots + (B_t)\right\}$$

$$= (A_m) - \frac{(A_m)}{N} \times N$$

$$= 0 .$$

$\boxed{\text{OR}}$

$$\Bigg[\quad \sum_{j=1}^{t} \delta_{mj} = \sum_{j=1}^{t} \left\{ (A_m B_j) - \frac{(A_m)(B_j)}{N} \right\}$$

$$= \sum_{j=1}^{t} (A_m B_j) - \frac{(A_m)}{N} \sum_{j=1}^{t} (B_j)$$

$$= (A_m) - \frac{(A_m)}{N} . N .$$

$$= 0 \quad \Bigg] .$$

∴ For a particular n, $\sum_{i=1}^{s} \delta_{in} = 0 .$

Now, there are st δ-quantities. In virtue of the relationship we have just proved, for any particular m only (t−1) of the t quantities δ_{mn} are independent. Similarly, for any n only (s−1) are independent. Hence, the total number of independent δ's is (s−1)(t−1).

Some Measur es Of Association :→ (c.v.)

(1) Mean Sauare Contingency :→ We may define a measure of association in terms of so called 'sauare contiengency'

$$\chi^2 = \sum_{m=1}^{s}\sum_{n=1}^{t}\left(\frac{\delta^2_{mn}}{(A_mB_n)_0}\right)$$

and call χ^2 the "sauare contingency".

We then define, $\phi^2 = \frac{\chi^2}{N}$

and call ϕ^2 the "mean-sauare contingency".

Clearly, χ^2 and ϕ^2, being the sums of squares, can't be negative. Then vanish if, and only if, every δ-number vanishes, in which case A and B are independent.

& ϕ^2 is not suitable for Comparative study.

$$\chi^2 = \sum_{m=1}^{s}\sum_{n=1}^{t}\frac{\delta^2_{mn}}{(A_mB_n)_0}$$

$$= N\left\{\sum_{m=1}^{s}\sum_{n=1}^{t}\frac{(A_mB_n)^2}{(A_m)(B_n)} - 1\right\}$$

$$= N\left\{\sum_{m=1}^{t}\frac{(A_mB_m)^2}{(A_m)(B_m)} - 1\right\} \quad [\text{where } s=t]$$

$$= N(t-1)$$

$\Leftrightarrow \phi^2 = t-1$.. Hence the limits of ϕ^2 vary in different systems.

~~(2) Karl Pearson's coefficient of mean-sauare contingency~~

Properties :→ (a) ϕ^2 is non-negative since it is the sum of sauares and ϕ^2 can take any non-negative any non-negative value, i.e. $0 \le \phi^2 \le \infty$.

(b) $\phi^2 = 0$ iff $\delta_{mn} = 0 \ \forall \ m, n$.

iff A and B are independent.

Limitations :→ (a) Since χ^2 can take any non-negative value then the limits of ϕ^2 vary in different situations. Hence ϕ^2 is not suitable to form a coefficient of association.

(b) In $t \times t$ table, in case of complete association so that $(A_m) = (B_m) = (A_mB_m) \ \forall \ m$ & $(A_mB_n) = 0 \ \forall \ m \ne n$, i.e. when only the leading diagonal freauencies are non-zero.

Why? $$\sum_{m=1}^{s}\sum_{n=1}^{t}\frac{\delta^2_{mn}}{(A_mB_n)_0}$$

$$=\sum_{m=1}^{s}\sum_{n=1}^{t}\frac{N}{(A_m)(B_n)}\left[(A_mB_n)-\frac{(A_m)(B_n)}{N}\right]^2$$

$$=N\sum_{m=1}^{s}\sum_{n=1}^{t}\frac{\left\{(A_mB_n)^2-\frac{2(A_mB_n)(A_m)(B_n)}{N}+\frac{(A_m)^2(B_n)^2}{N^2}\right\}}{(A_m)(B_n)}$$

$$=N\sum_{m=1}^{s}\sum_{n=1}^{t}\left\{(A_mB_n)^2+\frac{(A_m)(B_n)}{N^2}\Big((A_m)(B_n)-2(A_mB_n)N\Big)\right\}\Big/(A_m)(B_n)$$

$$=N\left[\sum_{m=1}^{s}\sum_{n=1}^{t}\left\{\frac{(A_mB_n)^2}{(A_m)(B_n)}\right\}+\frac{1}{N^2}\sum_{m=1}^{s}\sum_{n=1}^{t}\Big\{(A_m)(B_n)-2(A_mB_n)N\Big\}\right]$$

$$=N\left[\sum_{m=1}^{s}\sum_{n=1}^{t}\frac{(A_mB_n)^2}{(A_m)(B_n)}-1\right]$$

C.U.

(2) Karl Pearson's coefficient of mean-square contingency:—

The quantity ϕ^2 is not quite suitable in itself to form a coefficient, because its limits vary in different cases. The upper-limit is infinite as N increases. Pearson, therefore, proposed the coefficient C, defined by,

$$C=\sqrt{\frac{\chi^2}{N+\chi^2}}=\sqrt{\frac{\phi^2}{1+\phi^2}}.$$

This is called the "coefficient of mean-square contingency". In general, no sign should be attached to the root, for the coefficient merely shows whether two characters s are or are not independent; but in certain cases a conventional sign may be used. As for example — slight pigmentation of eyes and hair appear to go together, and the contingency may be regarded as positive. If slight pigmentation of eyes had been associated with marked pigmentation of hair, the contingency might have been regarded as negative.

c.u Properties :→

(1) C vanishes iff $\delta_{mn}=0 \ \forall (m,n)$
i.e. C = 0 iff A and B are independent of each other.

(2) the coefficient has one serious disadvantages.

c.u $\therefore 0 \le C < 1$ since $0 \le \chi^2 < N+\chi^2$

Limitations :→

(1) The coefficient C has one serious drawback that it never reaches the value 1 even in the case of perfect association.

(2) In t×t table, in case of complete association,

$\chi^2 = N(t-1)$ and

$$C = \sqrt{\frac{N(t-1)}{N+N(t-1)}} = \sqrt{\frac{t-1}{t}}$$

Note that, no greater association than the case

$$(A_m B_m) = (A_m) = (B_m) \ \forall m \text{, and}$$

$$(A_m B_n) = 0 \ \forall m \ne n$$

can be imagined in t×t table. Hence, $\chi^2 \le N(t-1)$ and $C \le \sqrt{\frac{t-1}{t}}$. The upper limits of C vary for different systems and hence it is not suitable for comparative study.

c.u (3) Tschuprow's Coefficient :→ Even in case of complete association the value of C is not unity and its maximum depends on the number of rows and columns of the table. To remedy the defect to which we have reffered above, Tschuprow proposed an alternative coefficient T; defined by,

$$T^2 = \left\{\frac{\chi^2}{N\sqrt{(s-1)(t-1)}}\right\}$$

$$= \frac{\phi^2}{\sqrt{(s-1)(t-1)}}$$

T = 1, in case of complete association in a t×t table but it will not be true in case of an s×t classification where $s \ne t$.

The coefficient varies between 0 and 1.

(4) Cramer's V^2 :

In an $s\times t$ table, if $s>t$ and $(A_mB_m) = (A_m)(B_m)$, $m = 1(1)t$; then $\chi^2 = N(t-1)$ but if $s<t$ & $(A_mB_m) = (A_m)(B_m)$, $m=1(1)t$; then $\chi^2 = N(s-1)$.

Hence for any $s\times t$ table, $\chi^2 \le N\min\{s-1, t-1\}$.

As Cramer pointed out, we may avoid these difficulties by defining a new measure:

$$V^2 = \frac{\chi^2}{N\min\{s-1, t-1\}}$$

Clearly V^2 lies between 0 and 1.

When, $s=t=2$, $V^2 = \frac{\chi^2}{N} = \phi^2$

Evidently, $V^2 = T^2$ when the table is square, otherwise V^2 exceeds T^2. Although the difference won't be very large unless s and t varies widely.

We also see that —

$$\frac{C^2}{T^2} = \frac{\sqrt{(s-1)(t-1)}}{1+\phi^2}$$

Remark:→ These measures of association, discussed above, are applicable to both nominal and ordinal data.

ORDINAL TRENDS :— Let us consider two jointly distributed attributes A and B. Further assume that the level of each can be arranged according to their degree of possession. Here it may happen that — as the responses on the attribute A increase towards its higher level, responses on B also increase towards its higher level or the responses on B decrease towards its lower level. Such type of association is often reffered to as a monotone trend association.

Concordance & Discordance :— A pair is said to be concordant pair if the subject ranked higher on A also ranked higher on B. A pair is said to be discordant pair if the subject ranking higher on A, ranks lower on B. A pair is said to be tied if the subjects have same classification on A and/or on B.

☑ Ordinal measures of association :—

Here we shall consider some measures based on the no. of concordance and the no. of discordance pairs.

(1) Goodman-Kruskal Gamma :— Let C & D be respectively the no. of concordance pairs and the no. of discordance pairs. If the pairs are united on both variables, then $\frac{C}{C+D}$ is the proportion of concordance pairs and $\frac{D}{C+D}$ is the proportion of discordance pairs. The difference $\left(\frac{C}{C+D}-\frac{D}{C+D}\right) = \left(\frac{C-D}{C+D}\right)$ is termed as Goodman-Kruskal Gamma (γ).

i.e. $\gamma = \frac{C-D}{C+D}$.

If $\gamma \gtrless 0$, i.e. $\frac{C}{C+D} \gtrless \frac{D}{C+D}$, then the majority of the pairs are in concordance (or discordance), i.e. characters A and B have the +ve (or, −ve) association.

Properties :—

(a) The Goodman-Kruskal Gamma (γ) treats the characters symmetrically — any one of them can be taken as response variable.

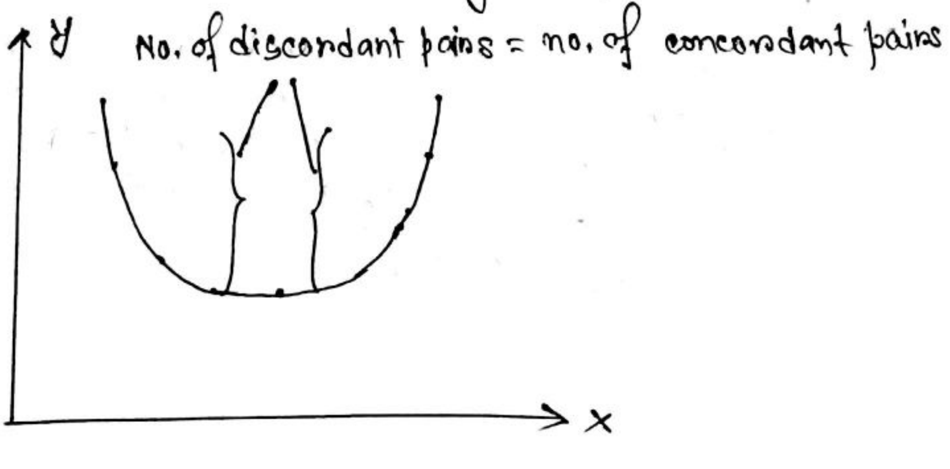
(b) Clearly, $|\gamma| \leq 1$, i.e. $-1 \leq \gamma \leq 1$, since $C \geq 0$ and $D \geq 0$.

(c) Under independence γ vanishes but the converse is not true. i.e. Independence $\Rightarrow \gamma = 0$ but the converse is not necessarily true. For a U-shaped joint-distn., one can have $C = D$., i.e. $\gamma = 0$ and the characters are not independent (pic:(c))

(d) $\gamma = \pm 1$ iff the characters are monotonically related.

$\gamma = 1 \Leftrightarrow D = 0$ iff the characters have monotonic increasing relation.

$\gamma = -1 \Leftrightarrow C = 0$ iff the characters have monotonic decreasing relation.



pic:(c)

REMARK :→ Note that, —

$$C = 2\sum_i\sum_j n_{ij}\left(\sum_{p>i}\sum_{k>j} n_{pk}\right)$$

$$D = 2\sum_i\sum_j n_{ij}\left(\sum_{p>i}\sum_{k<j} n_{pk}\right)$$

Let, T_A = No of tied pairs on the characters A

$$= \binom{n_{10}}{2} + \binom{n_{20}}{2} + \cdots + \binom{n_{s0}}{2}$$

$$= \sum_{i=1}^{s} \frac{n_{i0}(n_{i0}-1)}{2}$$, since n_{i0} members have the same characters A.

Similarly, T_B = No. of tied pairs on the characters B.

$$= \binom{n_{01}}{2} + \binom{n_{02}}{2} + \binom{n_{03}}{2} + \cdots + \binom{n_{0t}}{2}$$

$$= \sum_{j=1}^{t} \frac{n_{0j}(n_{0j}-1)}{2}$$

Now, T_{AB} = No. of tied pairs on both A and B

$$= \sum_{i=1}^{s}\sum_{j=1}^{t}\binom{n_{ij}}{2}$$

$$= \sum_i\sum_j \frac{n_{ij}(n_{ij}-1)}{2}$$

Note that → $\binom{N}{2} = C + D + T_A + T_B - T_{AB}$

(2) Kendall's τ :— Let (x_i, y_i) and (x_j, y_j) be two pairs of observation on two characters for a pair (i, j), $i<j$ of individuals.

Define, $a_{ij} = \text{sign}(x_i - x_j)$

$$= \begin{cases} 1 & , \text{if } x_i > x_j \\ 0 & , \text{if } x_i = x_j \\ -1 & \text{if } x_i < x_j \end{cases}$$

and, $b_{ij} = \text{sign}(y_i - y_j)$

Then, our measure of rank correlation will be based on the sum, $S = \sum\sum_{i<j} a_{ij} b_{ij}$

$= C - D$, which is the total score.

Here, we have two choices to make a measure of association, if we wish to standardize S to lie in the range $[-1, 1]$ and attain ± 1 in the extreme cases of complete disassociation and complete association.

(a) Kendall's τ_a :— If there are no ties, no a_{ij} or b_{ij} are zero. In that case, S coould vary in between $-\binom{N}{2}$ & $\binom{N}{2}$.

i.e. $-\binom{N}{2} \leq S \leq \binom{N}{2}$. Hence, we define,

$$\text{Kendall's } \tau_a = \frac{S}{\binom{N}{2}}$$, as a coefficient of association.

Clearly, whether τ_a attains the end points or not, completely depends on the no. of zero scores.

If some scores a_{ij} or b_{ij} are zero, the τ_a can no longer attain ±1.

[$$S = \sum\sum_{i<j} a_{ij} b_{ij} = C - D$$

If there are no ties, then $S = \pm\binom{N}{2}$ if all pairs are concordant or discordant.

If there are some ties, then —

$$-\binom{N}{2} < S < \binom{N}{2}$$ &

consequently τ_a can't attain ±1.]

Note that, $\boxed{\tau_a = \frac{S}{\binom{N}{2}} = \frac{C-D}{\binom{N}{2}}}$, which is same as the kendall's τ_a.

(b) Kendall's τ_b :— The correlation coefficient between two set of scores (a_{ij}) and (b_{ij}) is defined as —

$$\tau_b = \frac{\sum\sum_{i<j} a_{ij} b_{ij}}{\sqrt{\sum\sum_{i<j} a_{ij}^2}\sqrt{\sum\sum_{i<j} b_{ij}^2}}$$

In the presence of tied pairs, the measure τ_b will be of the form,

$$\tau_b = \frac{C - D}{\sqrt{\binom{N}{2} - T_A}\sqrt{\binom{N}{2} - T_B}}$$

[Now, $\sum\sum_{i<j} a_{ij}^2$ = No. of pairs (i,j), $i<j$ for which $\{x_i > x_j\}$ or $\{x_i < x_j\}$

$= \binom{N}{2}$ − No. of tied pairs on A

$= \binom{N}{2} - T_A$, where $T_A = \sum_{i=1}^{s} \frac{n_{i0}(n_{i0}-1)}{2}$

& Similarly, $\sum\sum_{i<j} b_{ij}^2 = \binom{N}{2} - T_B$.]

If there are no ties, then $\tau_a = \tau_b$, but if there are some ties then $\tau_b > \tau_a$.

Remark :→ The Goodman - kruskal γ can also be expressed as —

$$\gamma = \frac{C-D}{C+D} = \frac{C-D}{\binom{N}{2} - T_A - T_B + T_{AB}}.$$

It is important to note that γ, τ_a, τ_b all attain ±1 if all the observations are in the cells along the longest diagonal of the table.

(3) **Somer's d** :⟶ Somer proposed a measure of association for ordinal data, define,

$$d = \frac{C-D}{\binom{N}{2} - T_A} \text{ or } \frac{C-D}{\binom{N}{2} - T_B}$$ → as a coefficient of association.

In case of complete association, $(A_mB_m) = (A_m) = (B_m)$, $m = 1(1)\min(s,t)$, and all other cell frequencies are zero, i.e. all the observations lie in a longest diagonal of a table. Then

In that case,

$$S = \sum_{i<j}\sum a_{ij}b_{ij} = \left\{ \text{No. of pairs } (i,j), i<j \text{ for which } \underset{\text{or} >}{x_i < x_j} \left(\text{or } \underset{\text{or} >}{y_i < y_j}\right) \right\}$$

$$= \{\text{No of pairs united on } A \text{ (or } B)\}.$$

$$= \left\{\binom{N}{2} - T_A\right\} \text{ or } \left\{\binom{N}{2} - T_B\right\}.$$

Hence, d = 1.

But if all the observations lie on the longest diagonal of a table such that $(A_m B_{\overline{s-m+1}}) = (A_m) = (B_{\overline{s-m+1}})$ $\forall\, m = 1(1)\min(s,t)$ (assuming that $s<t$), and all other frequencies are zero.

Then,
$$S = \sum_{i<j}\sum a_{ij}b_{ij} = -\left\{ \text{No. of pairs } (i,j), i<j \text{ for which } x_i > x_j \text{ (or } y_i > y_j) \right\}$$

$$= -\left\{\binom{N}{2} - T_B\right\} \text{ or } -\left\{\binom{N}{2} - T_A\right\}$$

and then $d = -1$.

Remark :→ In the above situations, kendall's τ_b behaves similarly.

Logistic Regression :—

Consider a clinical trial where patients are given a treatment and response (=1 if success ; =0 if failure) is observed. In addition their age is also reported. Let y be the response variable and x be the age. Then we assumed

$$P(y=1) = \pi(x)$$

$$\pi(x) = \frac{e^{a+bx}}{1+e^{a+bx}} \text{ ; } a, b \text{ are unknown.}$$

Consider n points of data (x_i, y_i) ; $i=1(1)n$. then the corresponding regression is termed as "Logistic regression".

$$a+bx = \log \frac{\pi(x)}{1-\pi(x)} = \log \text{ odds ratio representation.}$$

Here, we can't use linear regression model i.e. $y=a+bx$, because for the unknown a and b may not take the value 0 or 1.

Ref :- Agresti :-

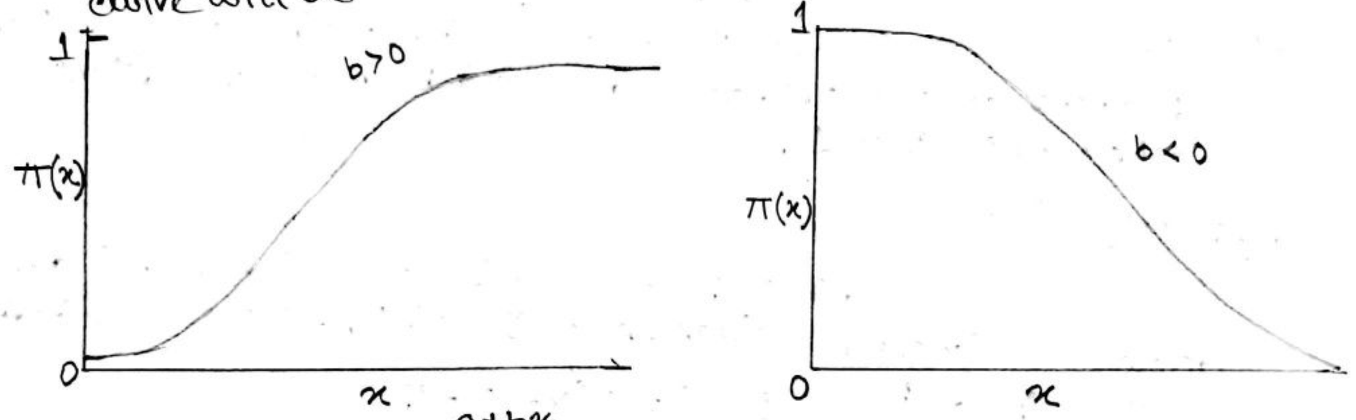
Let Y denote a response variable that can assume only two values, say 0 and 1. Denote the expected value of Y by $E(Y) = P(Y=1) = \pi$ and suppose that we want to model $\pi(x)$ on the values of explanatory variables $X=(x_1, x_2, \dots, x_k)$. The standard regression model has the form

$$\pi(x) = a + b_1x_1 + \dots + b_kx_k .$$

There are several difficulties with using ordinary least square principle to fit a model of standard regression conditions that make least squares estimates optimal are not there. For instance, the variance of Y is $\pi(x)(1-\pi(x))$, which is not constant over the range of values of the explanatory variables., standard distributional statements for estimators do not apply. Since Y is dichotomous rather than normally distributed.

A weighted least squares approach can be used to obtain more efficient estimates of the regression parameters in this model. The model itself is likely to be inaccurate in certain regions, however, if some x_i's are quantitative. They follows because the model predicts the impossible value $\pi < 0$ and $\pi > 1$ for sufficiently large or sufficiently small values of x_i. For a dichotomous response, $E(Y)$ can't be linearly related to x_i over an unbounded range of x_i values.

Logistic function :→ Because of the factors just discussed, it is often more appropriate to use a model that allows a curvilinear relationship between $E(Y)$ and each quantitative x_i. If we expect monotonic relationship then the regression curve will be



Here $\pi(x) = \frac{e^{a+bx}}{1+e^{a+bx}}$ called the logistic function. This function is monotonic with $\pi(x) \downarrow 0$ or $\pi(x) \uparrow 1$ as $x \uparrow \infty$ depending on where $b<0$ or $b>0$. Takes the value $\pi(x) = \frac{1}{2}$ at $x = -\frac{a}{b}$ and the curve has a steeper rate of increase around that value as $|b|$ increases. When $b>0$, it is the distribution function of the logistic random variable having mean $-\frac{a}{b}$ and s.d. $\frac{\pi}{\sqrt{3}b}$.

Here the odds of making response 1 instedd of response 0 is

$$\frac{\pi(x)}{(1-\pi(x))} = e^{a+bx}$$

The odds increases multiplicatively by e^b for every unit of x. The log odds has the simple linear relationship

$$\log\left[\frac{\pi(x)}{1-\pi(x)}\right] = a+bx$$

The model of the log odds is called logistic regression model. In this relationship the logit transformation yields a linear relationship for the logit model. When there are several explanatory variables, the logit model generalisis to

$$\log\left[\frac{\pi(x)}{1-\pi(x)}\right] = a+b_1x_1+\cdots+b_Kx_K.$$

Logit model:→ In applied mathematics & statistics, the logit of a number p between 0 and 1 is

$$\text{Logit}(P) = \log\left(\frac{P}{1-P}\right) = \log P - \log(1-P)$$

The logit function is the inverse of the sigonoid or logistic function. If p is a probability then $p/(1-p)$ is the correspondin odds and the logit of the probability is the logarithm of the odds.

Why do we use logistic regression rather than ordinary linear regression?

Ans:—

1) If we use linear regression, the predicted values will become greater than one and less than zero if you move far enough on the x-axis. Such values are theoritically inadmissible.

2) One of the assumption of regression is that variance of y is constant accross values of x. This can not be the case with a binary variable because the variance is $p(1-p)$. So, if p approaches to 1 or 0; variance approaches to zero.

3) The significance of testing of the weights b rest upon the assumption that erros of priction $(y-y')$ are normally distributed. Because y only takes values 0 and 1, this assumption is hard to justify. therefore, we can't use linear regression.

Relative Risk :— A value $\pi_1 - \pi_2$ of fixed size may have greater importance when both π_i are close to zero or 1 than when they are not. For a study comparing two treatments on the proportion of subjects who die, the difference between 0.010 and 0.001 and 0.410 and 0.401 are both 0.009. In such cases, the ratio of proportions is also informative. The relative risk is defined to be the ratio $\frac{\pi_1}{\pi_2}$. It can be any non negative real number. Relative risk of one corresponds to independence, e.g. for the proportions just given the relative risks are $\frac{0.010}{0.001} = 10$ and

$$\frac{0.410}{0.401} = 1.02 .$$

Comparing the rows on the 2nd response (failure) category gives a different relative risk, $\frac{1-\pi_1}{1-\pi_2}$.

2.v. **Odds Ratio :—** For a probability π of a success, the odds are define to be $\Omega = \frac{\pi}{1-\pi}$. The odds are non-negative with $\Omega > 1$ when a success is more likely than a failure, e.g.

$$\pi = 0.75 \Rightarrow \Omega = 3,$$ a success is thrice is as likely as a failure.

Inversely, $\pi = \frac{\Omega}{\Omega+1}$.

Refer again to 2X2 table, within row i, the odds of success instead of failure are $\Omega_i = \frac{\pi_i}{1-\pi_i}$.

The ratio of the odds Ω_1 and Ω_2 in the two rows,

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)},$$ is called the odds ratio.

For joint distn with cell probabilities $\{\pi_{ij}\}$, the equivalent defn of Ω_i could be

$$\Omega_i = \frac{\pi_{i1}}{\pi_{i2}} .$$ Hence,

$$\theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$$

Hence, odds ratio is also called cross-product ratio.

Properties of Odds ratio :— The odds ratio can be equal to any non-negative number. the condition $\Omega_1 = \Omega_2$ an hence (when all cell probabilities are +ve), $\theta = 1$ corresponds to independence of X and Y. When $1 < \theta < \infty$ subjects in row 1 are more likely to have a success than subjects in row 2; i.e. $\pi_1 > \pi_2$. For instance, when $\theta = 4$, the odds of success in row 1 are 4 times the odds in row 2, this does not mean that $\pi_1 = 4\pi_2$, i.e. relative risk of 4. When $0 < \theta < 1$, $\pi_1 < \pi_2$, one of the cell probabilities vanishes, θ becomes 0 or ∞.

Values of θ further from 1 in a given direction represents stronger association. Two values represent the same association, but in opposite directions, when one is the inverse of the others.

C.V.

Relationship between Odds ratio & Relative risk :—

$$\text{Odds ratio} = \text{relative risk} \times \frac{1-\pi_2}{1-\pi_1}.$$

Their magnitude are similiar whenever the probability π_i of the outcome of interest is close to zero for both groups thus when each π_i is small, the odds ratio provides a rough indication of the relative risk when it is not directly estimable, such as in case of controlled study.

We can compute the odds ratio, however, since it is determined by the conditional distributions in either direction. When the probability of the outcome of interest is very small, the population odds ratio and relative risk take similar values.

Logistic Regression :—

Let Y denote a binary response variable, for instance Y might indicate the choice of car (new or used) or the diagonasis of cancer (present or absent). Each observation has one of 2 outcomes denoted by 0 and 1. Now for a response variable Y and an explanatory variable X, consider the data.—

Setting of x	Values of Y	Totals
x_1	$y_{11}\ y_{12} \cdots y_{1N_1}$	$\sum_{j=1}^{N_1} y_{ij} = n_1$
x_2	$y_{21}\ y_{22} \cdots y_{2N_2}$	$\sum_{j=1}^{N_2} y_{2j} = n_2$
$\vdots$		$\vdots$
x_K	$y_{K1}\ y_{K2} \cdots y_{KN_K}$	$\sum_{j=1}^{N_K} y_{Kj} = n_K$

Here the true regression of y on x is given by the array mean:

$$\pi(x_i) = \bar{y}_i = \sum_{j=1}^{N_i} y_{ij} \Big/ N_i = \frac{n_i}{N_i}$$

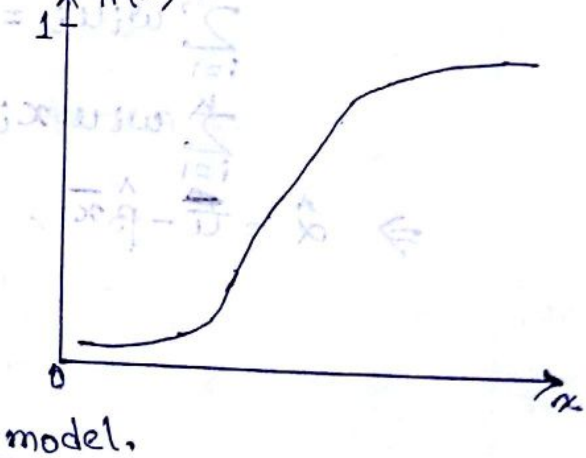
Since n_i is the number of y_{ij} which takes the value 1 when $x = x_i$ and $0 \le n_i \le N_i$. Hence, $0 \le \bar{y}_i \le 1$ and

$$0 < \frac{\bar{y}_i}{1-\bar{y}_i} < \infty \text{. Then, } -\infty < \log_e\left(\frac{\bar{y}_i}{1-\bar{y}_i}\right) < \infty .$$

i.e. $\log_e\left(\frac{\bar{y}_i}{1-\bar{y}_i}\right)$ can be any real number, the real nos are also the range of any linear predictor such as $(\alpha+\beta x)$. We can assume the regression model as $\log_e\left(\frac{\bar{y}_i}{1-\bar{y}_i}\right) = \alpha + \beta x_i$

$$\Leftrightarrow y_i = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}$$

[Usually binary data, result from a non-linear relationship between $\pi(x_i) = \bar{y}_i$ and x_i. A fixed change in x has less impact on $\pi(x)$ when $\pi(x)$ is near 0 or 1 than when $\pi(x)$ is near 0.5.]



In practice, non linear relationship between $\pi(x)$ and x is monotonic. Most important curve with the above shape is the logistic curve.

An appropriate regression model is

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}},$$

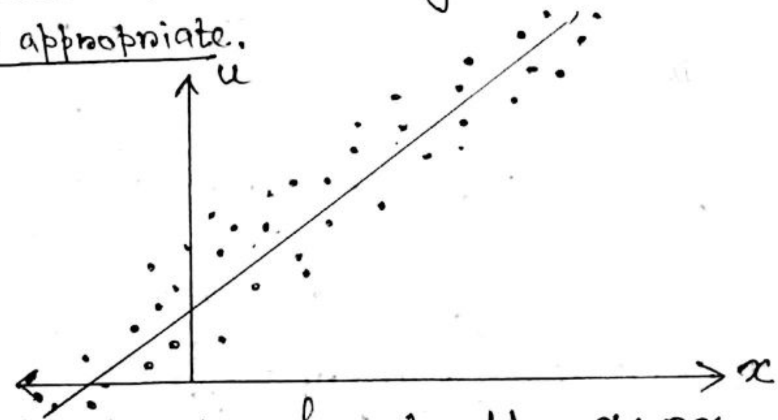
which is called logistic regression model.

C.U.

Looking at the data & fitting of logistic regression eauation

Before fitting logistic model, look at the data to check whether the logistic regression is appropriate or not. Since Y takes any of the values 0 and 1, it is difficult to check by plotting Y-values for different X.

If the quantities $u_i = \log\left(\frac{\bar{y}_i}{1-\bar{y}_i}\right) = \log\left(\frac{n_i}{N_i-n_i}\right)$ are plotted against x_i, $i=1(1)k$. and the points (x_i, u_i) are near about a line. then, by definition, the logistic regression is appropriate.



C.U.

- **Fitting :** Consider the predicting formula $U_x = \alpha + \beta x$. When $x = x_i$, then the predicting value of $u_i = \log\left(\frac{\bar{y}_i}{1-\bar{y}_i}\right)$ is $U_{x_i} = \alpha + \beta x_i$ and the corresponding error $e_i = u_i - U_{x_i} = (u_i - \alpha - \beta x_i)$.

To determine α and β, we shall use method of weighted least squares and we shall minimise

$$S^2 = \sum_{i=1}^{k} \omega_i e_i^2 = \sum_{i=1}^{k} \omega_i (u_i - \alpha - \beta x_i)^2,$$

where the weights $\omega_i = N_i \bar{y}_i (1-\bar{y}_i) = N_i \pi(x_i)(1-\pi(x_i))$, with respect to α and β.

The normal equations are:

$$\sum_{i=1}^{k} \omega_i u_i = \alpha \sum_{i=1}^{k} \omega_i + \beta \sum_{i=1}^{k} \omega_i x_i$$

$$\sum_{i=1}^{k} \omega_i u_i x_i = \alpha \sum_{i=1}^{k} \omega_i x_i + \beta \sum_{i} \omega_i x_i^2$$

$$\Rightarrow \hat{\alpha} = \bar{u} - \hat{\beta}\bar{x}, \quad \hat{\beta} = \frac{\sum_{i=1}^{k} \omega_i u_i x_i - N\bar{u}\bar{x}}{\sum \omega_i x_i^2 - N\bar{x}^2},$$ where

$$\bar{u} = \frac{\sum \omega_i u_i}{\sum \omega_i} \text{ and } \bar{x} = \frac{\sum \omega_i x_i}{\sum \omega_i}.$$

Question:— What is the rationale behind the Yule's coefficient of association ?

Soln.→ δ is a measure of independence of A and B. To get a normed measure, we divide $N\delta$, the difference between $(AB)(\alpha\beta)$ & $(A\beta)(\alpha B)$ by the sum $(AB)(\alpha\beta)+(A\beta+\alpha B)$ and the resulting coefficient is $Q = \dfrac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$, which is suitable for comparison of two or more data sets.

Question :— What is the rationale behind Yule's coefficient of colligation ?

Soln.→ Empirical study reveals that the coefficient Q overestimate the extent of association. Hence, Yule provides a measure by considering lower orders of the cell frequencies as Y, a coefficient of colligation. Note that $Q = \dfrac{2Y}{1+Y^2} \geq Y$.

Question:— Let $x_i = \begin{cases} 1, & i\text{th individual possess A and similarly} \\ 0, & i\text{th individual possess } \alpha. \end{cases}$

$y_i = \begin{cases} 1, & i\text{th individual possess B} \\ 0, & i\text{th individual possess } \beta. \end{cases}$

Find the correlation coefficient between x and y based on the data $\{(x_i, y_i) : i = 1(1)n\}$.

Soln.→

$$r_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)^{1/2}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)^{1/2}}$$

$$= \frac{f_{11} - n \cdot \frac{f_{10}}{n} \cdot \frac{f_{01}}{n}}{\left\{f_{10} - \frac{f_{10}^2}{n}\right\}^{1/2} \left\{f_{01} - \frac{f_{01}^2}{n}\right\}^{1/2}}$$

$$= \frac{(f_{11}f_{22} - f_{12}f_{21})/n}{\left(\frac{f_{10}f_{20}}{n}\right)^{1/2}\left(\frac{f_{01}f_{02}}{n}\right)^{1/2}}$$

$$= \frac{f_{11}f_{22} - f_{12}f_{21}}{\sqrt{f_{10}f_{20}\,f_{01}f_{02}}}$$

$$= \frac{n\delta}{\sqrt{f_{10}f_{20}f_{01}f_{02}}}$$

$y \backslash x$	1	0	
1	f_{11}	f_{21}	f_{01}
0	f_{12}	f_{22}	f_{02}
	f_{10}	f_{20}	n

Hence, $r_{xy} = 0$ iff $\delta = 0$ iff A & B are independent.

Problem: For a 2x2 table, Yule introduces

$$Q = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}}$$

i) S.T. in a 2x2 table, Goodman-kruskal's $\gamma = Q$.

ii) S.T. the Yule coefficient can be written as

$Q = \dfrac{\hat{\theta}-1}{\hat{\theta}+1}$, a monotonic transformation of $\hat{\theta}$ from $[0,\infty)$ to $[-1,1]$.

Soln.→ i)

$$\gamma = \frac{C-D}{C+D}$$

Y \ X	Y_1	Y_2
x_1	n_{11}	n_{12}
X_2	n_{21}	n_{22}

Here, C = No. of pairs (i,j), $i<j$ for which

$\{x_i > x_j, Y_i > Y_i\}$ or $\{X_i < x_j, Y_i < Y_j\}$

$= n_{11} \times n_{22}$

and D = No. of pairs (i,j), $i<j$ for which $\{X_i > X_j, Y_i < Y_j\}$ or $\{X_i < X_j, Y_i > Y_j\}$

$= n_{21} \times n_{12}$

$$\therefore \gamma = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}} = Q$$

ii) $\hat{\theta}$ = the sample odds ratio or α-measure

$$= \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

$$\therefore Q = \frac{\hat{\theta}-1}{\hat{\theta}+1},$$

$$\frac{dQ}{d\hat{\theta}} = \frac{d}{d\hat{\theta}}\left(1 - \frac{2}{\hat{\theta}+1}\right) = \frac{2}{(\hat{\theta}+1)^2} > 0$$

$\Rightarrow Q$ is monotonic increasing in $\hat{\theta}$.

As $0 \le \hat{\theta} < \infty$, $1 \le \hat{\theta}+1 < \infty$

$$\Rightarrow 0 < \frac{2}{\hat{\theta}+1} \le 2$$

$$\Rightarrow -1 \le 1 - \frac{2}{\hat{\theta}+1} \le 1$$

$$\Rightarrow -1 \le Q \le 1.$$

Categorical Data Analysis

Categorical Table :— Let us consider two categorical variable A and B, where A takes values on p categories, say, $A_1, A_2, \dots, A_p$ and B takes values $B_1, B_2, \dots, B_q$, on q categories.

Let us define the probability,

$$\pi_{ij} = P(A = A_i \cap B = B_j) \quad \text{for } i = 1(1)p,\ j = 1(1)q.$$

If we construct a table consisting of p rows for the p-categories of A and q columns for the q categories of B, then we get a p×q table as follows:

A \ B	B_1	B_2	$\cdots$	B_j	$\cdots$	B_q	
A_1	π_{11}	π_{12}	$\cdots$	π_{1j}	$\cdots$	π_{1q}	π_{10}
A_2	π_{21}	π_{22}	$\cdots$	π_{2j}	$\cdots$	π_{2q}	π_{20}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_i	π_{i1}	π_{i2}	$\cdots$	π_{ij}	$\cdots$	π_{iq}	π_{i0}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_p	π_{p1}	π_{p2}	$\cdots$	π_{pj}	$\cdots$	π_{pq}	π_{p0}
	π_{01}	π_{02}	$\cdots$	π_{0j}	$\cdots$	π_{0q}	

This table is called a p× q contingency table. Here —

$$\pi_{i0} = \sum_{j=1}^{q} \pi_{ij} \quad \& \quad \pi_{0j} = \sum_{i=1}^{p} \pi_{ij}$$

are called Marginal Probabilities.

π_{ij}'s are called the Joint probabilities and with their probabilities we can define the conditional probabilities in this table as follows:

$$\pi_{j|i} = P(B = B_j \mid A = A_i) = \frac{P(A = A_i \cap B = B_j)}{P(A = A_i)} = \frac{\pi_{ij}}{\pi_{i0}}$$

Independence of two categorical variables: —

Two categorical variables A and B taking values on two categorical scales consisting of the categories $A_1, A_2, \ldots, A_p$ and $B_1, B_2, \ldots, B_q$, respectively, are independent if the conditional probability of assuming a value by one variable, say B, remains the same for any value, assumed by the other variable.

For every j, $\pi_{j|i} = P(B=B_j \mid A=A_i) = \pi_{oj}$ $\forall i$

i.e. for every j, $\dfrac{\pi_{ij}}{\pi_{io}} = \pi_{oj}$ $\forall i$

$$\Rightarrow \pi_{ij} = \pi_{io}\cdot\pi_{oj} \quad \forall i, j$$

Difference measures of independence in a 2×2 contingency table: —

Let there be two groups of subjects and let each of them be allowed to response on a binary variable. Let the categorical variable corresponding to the group of subject denoted by X and that corresponding to the response be denoted by Y. Then both X and Y are binary variables.

Response / Groups	Yes	No
Gr-I	π_1	$1-\pi_1$
Gr-II	π_2	$1-\pi_2$

In the adjacent table, let π_i $(i=1,2)$ be the proportion of 'Yes' response in the i-th group. Note that, here the probabilities π_1 and π_2 are actually conditional probability given for the group.

Now if the categorical variable X and Y are independent, Then $\pi_1 - \pi_2 = 0$. Now considering another table —

Response / Groups	Yes	No	No. of subjects
Gr-I	x_1	$n_1 - x_1$	n_1
Gr-II	x_2	$n_2 - x_2$	n_2

Now, we can judge independence from the difference of the observed proportion

$p_1 = \dfrac{x_1}{n_1}$ and $p_2 = \dfrac{x_2}{n_2}$.

If $p_1 - p_2$ is close to zero then we can predict independence, otherwise, association is believed to exist between the categorical variable X and Y.

Limitations:→ Limitation of this measures of independence is that when π is close to 0 or 1, the difference may lead to a wrong decision about independence; for instance, let $\pi_1 = 0.010$ and $\pi_2 = 0.001$, then $\pi_1 - \pi_2 = 0.009$, again let $\pi_1^* = 0.410$ and $\pi_2^* = 0.401$, then $\pi_1^* - \pi_2^* = 0.009$, i.e. difference of probabilities are same in both the cases, though $\frac{\pi_1}{\pi_2} = 10$ and $\frac{\pi_1^*}{\pi_2^*} = 1.2$. Limits of this measure are -1 and $+1$, $\pi_1 - \pi_2 = +1 \Rightarrow$ extreme +ve association, $\pi_1 - \pi_2 = -1 \Rightarrow$ extreme −ve association.

Example:-

Groups	AGES	YES	NO	TOTAL
Gr-I	Age ≥ 70	60	0	$60 = n_1$
Gr-II	Age ≤ 20	0	40	$40 = n_2$

$\pi_1 = 1;\ \pi_2 = 0$

$\therefore \boxed{\pi_1 - \pi_2 = 1}$

$\pi_1 = 0;\ \pi_2 = 1$

$\therefore \boxed{\pi_1 - \pi_2 = -1}$

where, $\pi_1 = \frac{x_1}{n_1}$, $\pi_2 = \frac{x_2}{n_2}$.

ODDS RATIO:- Let us compare two groups of subjects on a binary response variable Y. Let the categorical variable Y be allowed to take values 1 and 2. Then we denote by X another categorical variable, also taking values 1 and 2, to denote the two groups of subjects:

Groups (X)	Response (Y) 1	2	
1	π_1	$1-\pi_1$	1
2	π_2	$1-\pi_2$	1

If, we consider the response '1' as "success" with probability π, then the odds of success is defined by $\frac{\pi}{1-\pi}$. In the present case the odds of success for the 1st group is

$\text{Odds}_1 = \frac{\pi_1}{1-\pi_1}$ and that for the 2nd group is

$\text{Odds}_2 = \frac{\pi_2}{1-\pi_2}$.

Then, we define the Odds Ratio of the two groups by
$$\theta = \frac{\text{Odd } S_1}{\text{Odd } S_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}.$$
when two categorical variables X and Y are independent, $\pi_1 = \pi_2$, then the odds ratio$(\theta) = 1$.
When $\theta > 1$, we say that X and Y are +vely associated.
& $\theta < 1$, we say that X and Y are -vely associated.

PROPERTIES OF ODDS RATIO :-

1) Odds Ratio can take any non-negative value.

2) $\theta = 1 \Rightarrow$ Independence.
$\theta > 1 \Rightarrow$ positive association.
$\theta < 1 \Rightarrow$ negative association.

3) If θ and θ^* are two odds ratio s such that
$\theta = \frac{1}{\theta^*}$; then we can calculate that degree of association is the same in both the cases, though the direction of association is opposite.

4) For a table with joint probabilities, the odds ratio is given by
$$\theta = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

Proof:- In the table, we can write the odds ratio in terms of conditional probabilities
$$\pi_{1|1} = \frac{\pi_{11}}{\pi_{11}+\pi_{12}} \text{ and } \pi_{1|2} = \frac{\pi_{21}}{\pi_{21}+\pi_{22}}$$
Then,
$$\theta = \frac{\pi_{1|1}/(1-\pi_{1|1})}{\pi_{1|2}/(1-\pi_{1|2})} = \frac{\pi_{1|1}(1-\pi_{1|2})}{\pi_{1|2}(1-\pi_{1|1})}$$
$$= \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

5) If any row or any column of the table is interchanged, then the odds ratio is reversed.

Proof:-

	1	2
1	π_1	$1-\pi_1$
2	π_2	$1-\pi_2$

Odds ratio is

$$\theta = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$$

Now, if the 1st and 2nd rows of the table are interchanged to get the new table.

	1	2
1	π_2	$1-\pi_2$
2	π_1	$1-\pi_1$

Then the odds ratio is

$$\theta^* = \frac{\pi_2(1-\pi_1)}{\pi_1(1-\pi_2)}$$

$$= \frac{1}{\theta}.$$

6. If the orientation of the table is reversed then the odds ratio does not change.

Proof:→

X \ Y	1	2
1	π_1	$1-\pi_1$
2	π_2	$1-\pi_2$

Here the odds ratio is

$$\theta = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$$

Now, if the orientation of this table is reversed to get the new table.

X \ Y	1	2
1	π_1	π_2
2	$1-\pi_1$	$1-\pi_2$

Here, the odds ratio is

$$\theta^* = \frac{\pi_1/\pi_2}{(1-\pi_1)/(1-\pi_2)}$$

$$= \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$$

$$= \theta.$$

c.u. **PROBLEM :→** Show that sample odds ratio does not change when both cell counts within any row are multiplied by a non-zero constant or when both cell counts within any column are multiplied by a non-zero constant. Discuss the implication of the above result using a real life example.

ANS:→

	1	2
1	π_1	$1-\pi_1$
2	π_2	$1-\pi_2$

Odds ratio is

$$\theta = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$$

Now, if the both cell counts with in any row are multiplied by a constant (nonzero), say, k, we get the new table.

	1	2
1	$k\pi_1$	$k(1-\pi_1)$
2	π_2	$1-\pi_2$

Then the odds ratio is

$$\theta^* = \frac{k\pi_1(1-\pi_2)}{\pi_2(k(1-\pi_1)}$$
$$= \theta$$

So, odds ratio does not change.

An implication of the multiplicative invariance property is that the sample odds ratio estimates the same characteristic (θ) even when we select disproportionately large or small samples from marginal categories of a variable. For instance, suppose a study investigates the association between vaccination and catching a certain strain of flu. For a retrospective design, the sample odds ratio estimates the same characteristic whether we randomly sample (1) 100 people who got the flu and 100 people who didn't, (2) 150 people who got the flu and 50 people who didn't, in each case classifying subjects on whether they took the vaccine. In fact, the odds ratio is equally valid for retrospective, prospective, or cross-sectional sampling designs. We would estimate the same characteristic if (3) we randomly sample 100 people who took the vaccine and 100 people who didn't, and then classify them on whether they got the flu, or (4) we randomly sample 200 people and classify them on whether they took the vaccine and whether they got the flu.

—— × ——