



**U. P. Rajarshi Tandon Open University,
Prayagraj**

**Master of Science/
Master of Arts**

**PGMM-108N/MAMM-108N
Mathematical Statistics**

Course Design Committee

Prof. Ashutosh Gupta, Director, School of Sciences, UPRTOU, Prayagraj	Chairman
Prof. A. K. Malik School of Sciences, UPRTOU, Prayagraj	Coordinator
Prof. Mukesh Kumar Department of Mathematics, MNIT, Prayagraj	Member
Dr. A. K. Pandey Associate Professor, ECC, Prayagraj	Member
Dr. Raghvendra Singh Assistant Professor, Mathematics School of Sciences,UPRTOU, Prayagraj	Invited Member/Secretary

Course Preparation Committee

Prof. Ram Lal Department of Mathematics and Statistics, Allahabad Agricultural Institute (Deemed University), Naini, Allahabad	Writer/s	Block - 1
Dr. Sheela Mishra Department of Statistics, Lucknow University, Lucknow		
Prof. S. K. Pandey Department of Statistics, Lucknow University, Lucknow	Reviewer/s	
Prof. V. P. Ojha Department of Statistics and Mathematics, D. D. U., Gorakhpur University, Gorakhpur	Editor/s	Block - 2
Prof. S. K. Pandey Department of Statistics, Lucknow University, Lucknow		
Prof. S. K. Upadhyay Department of Statistics, Banaras Hindu University, Varanasi		
Dr. Shruti Assistant Professor, School of Sciences U.P. Rajarshi Tandon Open University, Allahabad	Writer/s	
Prof. G. S. Pandey Department of Statistics, Allahabad University, Allahabad	Reviewer/s	
Prof. V. P. Ojha Department of Statistics and Mathematics, D. D. U., Gorakhpur University, Gorakhpur	Editor/s	
Prof. S. K. Pandey Department of Statistics, Lucknow University, Lucknow		
Prof. S. K. Upadhyay Department of Statistics, Banaras Hindu University, Varanasi		

Prof. B. P. Singh Department of Statistics, Banaras Hindu University, Varanasi Dr. Alok Kumar Department of Community Medicine, Institute of Medical Sciences, Banaras Hindu University, Varanasi	Writer/s	Block - 3
Dr. Sanjeeva Kumar Department of Statistics, Banaras Hindu University, Varanasi	Reviewer/s	
Prof. R. C. Yadava Department of Statistics, Banaras Hindu University, Varanasi Prof. V. P. Ojha Department of Statistics and Mathematics, D. D. U., Gorakhpur University, Gorakhpur Prof. S. K. Pandey Department of Statistics, Lucknow University, Lucknow	Editor/s	
Prof. B. P. Singh Department of Statistics, Banaras Hindu University, Varanasi Dr. Sheela Mishra Department of Statistics, Lucknow University, Lucknow Dr. Alok Kumar Department of Community Medicine, Institute of Medical Sciences, Banaras Hindu University, Varanasi Dr. Shruti Assistant Professor, School of Sciences U.P. Rajarshi Tandon Open University, Allahabad	Writer/s	Block - 4
Prof. V. P. Ojha Department of Statistics and Mathematics, D. D. U., Gorakhpur University, Gorakhpur Prof. S. K. Pandey Department of Statistics, Lucknow University, Lucknow Dr. Sanjeeva Kumar Department of Statistics, Banaras Hindu University, Varanasi	Reviewer/s	
Prof. V. P. Ojha Department of Statistics and Mathematics, D. D. U., Gorakhpur University, Gorakhpur Prof. R. C. Yadava Department of Statistics, Banaras Hindu University, Varanasi	Editor/s	

© UPRTOU, Prayagraj- 2024
PGMM –108N: MATHEMATICAL STATISTICS

ISBN-

©All Rights are reserved. No part of this work may be reproduced in any form, by mimeograph or any other means, without permission in writing from the Uttar Pradesh Rajarshi Tondon Open University, Prayagraj. Printed and Published by Vinay Kumar Singh, Registrar, Uttar Pradesh Rajarshi Tandon Open University, 2024.

Printed By : Chandrakala Universal Pvt. 42/7 Jawahar Lal Neharu Road, Prayagraj.

Syllabus

PGMM-108N/MAMM-108N: Mathematical Statistics

Block-1: Data Collection and its Representation

Unit-1: Data Collection and Tabulation

Meanings, Definitions and Applications of Statistics, Measurements and Scale, Measurements of qualitative data, Methods of data collection, Types of data.

Unit-2: Representation of Data- I (Diagrammatical representation)

Frequency distribution, Tabulation of data, Diagrammatical Representation of data, Bar diagram, Multiple bar diagram, Divided bar diagram, Percentage bar diagram, Pie chart, Pictogram, leaf chart.

Unit-3: Representation of Data- II (Graphical representation)

Graphical representation of frequency distribution, Histogram, Frequency polygon, Frequency curve, Ogive.

Block-2: Measures of Central Tendency and Dispersion

Unit-4: Measures of Central Tendency-I

Types of measures of central tendency, Arithmetic mean, Fundamental Theorems on Arithmetic mean, Geometric mean, Harmonic mean.

Unit-5: Measures of Central Tendency-II

Median, Mode, Percentiles, Deciles, and Quartiles.

Unit-6: Measures of Dispersion

Types of measures of Dispersion, Range, Mean Deviation, Variance and Standard deviation, Effect of change of origin and scale, Relationship between measures of central tendency and measures of dispersion, Coefficient of variation.

Block-3: Moments, Skewness and Kurtosis

Unit-7: Moments and Raw Moments

Definition of moments, raw moments for ungrouped data, raw moments for grouped data.

Unit-8: Central Moments

Central moments, Factorial moments, Interrelationship between various moments, effect of change of origin and scale on moments, Charlier's checks, Sheppard's correction for moments.

Unit-9: Skewness and Kurtosis

Definition of skewness, Measures of skewness, Pearson's coefficient, Bowley's coefficients, Kurtosis, Measures of Kurtosis, effect of change of origin and scale.

Block-4: Correlation and Regression

Unit-10: Bivariate Data and Correlation

Scatter Diagram, Karl Pearson's coefficient of correlation, Properties of correlation coefficient, limits of correlation coefficient, Effect of change of origin and scale on correlation coefficient.

Unit-11: Regressions

Regressions, linear regression model, principal of least square.

Unit-12: Line of Regressions

Regression lines, Regression coefficient, Properties of Regression coefficients.

Unit-13: Correlation and Intra Class Correlation

Rank correlation coefficient, Spearman's rank correlation coefficients, rank correlation coefficient for tied ranks, Intra-class correlation, some remarks on Intra-class correlation.

Unit-14: Theory of Attributes

Combinations, Classes and Class frequencies of Attributes, Dichotomous Classification, Consistency of data, joint distribution of attributes, Contingency tables, Independence and Association of Attributes, Measures of Association, Yates Correction.



U. P. Rajarshi Tandon
Open University

Master of Science

PGMM -108N

Mathematical Statistics

Block: 1 Data Collection and its Representation

Unit-1: Data Collection and Tabulation	13-28
Unit-2: Representation of Data- I (Diagrammatical representation)	29-54
Unit-3: Representation of Data- II (Graphical representation)	55-66

Block: 2 Measures of Central Tendency and Dispersion

Unit-4: Measures of Central Tendency-I	69-100
Unit-5: Measures of Central Tendency-II	101-122
Unit-6: Measures of Dispersion	123-148

Block: 3 Moments, Skewness and Kurtosis

Unit-7: Moments and Raw Moments	151-158
Unit-8: Central Moments	159-178
Unit-9: Skewness and Kurtosis	179-194

Block: 4 Correlation and Regression

Unit-10: Bivariate Data and Correlation	197-236
Unit-11: Regressions	237-252
Unit-12: Line of Regressions	253-282
Unit-13: Correlation and Intra Class Correlation	283-306
Unit-14: Theory of Attributes	307-332



U. P. Rajarshi Tandon
Open University

Master of Science

PGMM -108N

Mathematical Statistics

Block

1

Data Collection and its Representation

Unit- 1 Data Collection and Tabulation	13-28
Unit- 2 Representation of Data- I (Diagrammatical representation)	29-54
Unit- 3 Representation of Data- II (Graphical representation)	55-66

Block-1

Data Collection and its Representation

The present SLM on **Statistical Methods** consists of four Blocks. *Block - 1 – Data Collection and its Representation* has three units; *Block - 2 – Measures of Central Tendency and Dispersion* has three units; *Block - 3 – Moments, Skewness and Kurtosis* has three units and at the last *Block - 4 Correlation and Regression (Two Variables and Association)* has five units.

The **Block - 1 – Data Collection and its Representation** consist three units and deals with the introduction of Statistics and methods of data collection and also its representation. The *first unit* of this block introduce origin of Statistics, its meaning, definition and applications along with methods of data collection, measurements and scales. In the *second unit*, frequency distribution has been given. Here, Pie Chart, Bar Diagram, Pictograms and Leaf Chart have also been discussed. And the *third unit* deals with the graphical representation of data along with Histogram, Frequency Polygon, Frequency Curve and Ogives.

In the span started form the later 19th century till date Statistics have taken up unprecedented dimensions and now embrace almost every sphere of nature and human activity. Everyday statistical thinking is becoming more and more indispensable for an efficient citizenship. Through the comparative studies of the qualities and prices of the commodities, even a layman makes use of the statistical methods when as customer one decides as to what quality and from which dealer one should purchase one's daily provisions. There is no newspaper or a periodical these days without having a definite bearing upon statistics. Because of this rapid development and tremendous advancement in recent past, the elementary knowledge of Statistics has become a part of the general education in many advanced and developing countries these days. There is no ground for misgiving regarding practical realization of the dream of **H. H. G. Wells** "*statistical thinking one day be as necessary for efficient citizenship as the ability to read and write*".

At the end of every block/unit the summary, self-assessment questions and further readings are given.

Unit - 1: Data Collection and Tabulation Structure

Structure

- 1.1 Introduction**
- 1.2 Objectives**
- 1.3 Meaning and Definitions of Statistics**
- 1.4 Universe/Population**
- 1.5 Statistical Problems/ Limitations**
- 1.6 Measurements and Scales**
- 1.7 Measurement of Qualitative Data**
- 1.8 Measurement of Data Collection**
- 1.9 Summary**
- 1.10 Self-Assessment Questions**
- 1.11 Further Readings**

1.1 Introduction

The word “statistics” means status in Latin and “an organized political state” in German and may have been derived from either of them. Initially statistics was known as the science of statecraft and was used by the government to collect the various information needed to administer the state. The great philosopher Chanakya also recognized in his “Arthshastra” that for an efficient state management, the ruler should keep himself informed about the composition of the state population with respect to its various aspects such as: Literacy, public health, income and cost of living etc. In absence of these facts, (which were named statistics later on) the administration may become like groping in darkness.

1.2 Objectives

After going through this unit you should be able to

- Know origin of Statistics, its meaning, definitions and applications
- Define universe/Population
- Define statistical problems and Limitations
- Know measurements and scales
- Distinguish between discrete and continuous data
- Know methods of data collection, primary and secondary data

1.3 Meaning and Definitions of Statistics:

(a) Statistics and Statistical Data:

The word Statistics is used to convey two different senses and is defined differently in each case. One the plural of “Statistics” referring to the numerical data collected in an orderly manner with some specific objective in view. A. L. Bowley defines Statistics as numerical statements of facts in any department of enquiry placed in relation to each other. However, Prof. Hrace Secrist defined it as

“By Statistics we mean aggregate of facts affected to a marked extent by multiplicity of causes numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other.”

(b) Statistics as Statistical methods or as a tool of analysis:

When the word carries a singular sense it refers to the science of theory and techniques that are used to collect, represent, analyze and draw conclusions from the data. A. L. Bowley defined Statistics as the “science of measurements of social organism, regarded as a whole in all its manifestation”. In fact, a number of definitions of statistics denoting singularity are available but perhaps the best one available so far is given by Croxton and Cowden as

“Statistics may be defined as the science of collection, presentation, analysis and interpretation of numerical data.”

On the basis of these ideas, we can broadly **summarize that statistics is a science of**

- Collecting numerical information (data)
- Classification, summarization, organization and analysis of data
- Evaluation of the numerical information (data)
- Drawing conclusions based on evaluation of data

Application of Statistics

There has been a tremendous growth in the last century that Statistics keeps a roll to play in almost every branch of human knowledge; it may be the proper functioning of business and industry, understanding the principles of commerce and economics or the development of the various scientific theories and what not. A few of the multitudes of channels that confront statistics these days are as follows:

- **Statistics in business commerce and industry:** The important areas of business or industry are (a) Production (b) Marketing (c) Personnel (d) Finance and (e) Accounting, where the main functions of Statistics in a practical field of working are planning of operations, establishment of standards and their control. In business problems these statistical functions are conducted either in isolation or in mutual combinations. For example, Statistics is used for quality control in production and is employed for the analysis of sales and marketing in business. Wages and allowances of employees are fixed up on the basis of index numbers.

Statistical analysis of costing and accounting data is made for ascertaining profit or loss and knowing the financial position of the concern at a particular point. Statistical methods are very common and useful to accounts. Audits are done with speed and reliance through sampling. An estimate of the relationship between the cost and volume of production can be made through statistical studies of the past data.

- **Statistical methods** provide a valuable assistance for the study solution and formulation of **economic policies** on topics like production, distribution of wealth, demand and supply, etc. that no economist can afford to go without their exhaustive studies. **The government intervention in the national economy**, the growth of large scale entrepreneurial activity and introduction of scientific methods into various parts of business administration has stimulated and contributed to the rapid development of economic- statistics.

- A student of **Physics or Chemistry** or of any other **pure science**, while conducting an experiment in the laboratory has necessity to rely upon the application of statistics. An experiment is repeated, its reading varies and in order to reach closest to the accurate result, one has to tabulate them and an average is calculated. In fact, **higher studies in every science** need application of the statistical laws like correlation, regression, dispersion, approximation, probability and the tests of significance, etc.

- Innumerable illustrations can be given to show that in **biology** there are frequent applications of statistic. Tests of significance are applied to compare the effects of two or more drugs; the law of probability is employed in irradiation when the cells in the retina of eye are exposed to the light; chart is used to study heart beats through electrocardiograms, and the like. In **agriculture**; the comparison of varieties of seeds or of fertilizers is made through the principles of analysis of variance based on sampling theory.

The very fact that **industrial, medical, agricultural, bio statistics** and many more like that are now separate branch of study which speaks of every expanding scope of statistics and its indispensability in these areas. Statistics also provides a good device of saving time, material and personnel in different studies.

Statistical Applications may broadly be classified under following two disciplines;

(i) Descriptive Statistics:

In descriptive statistics we summarize or describe the data set at hand and evaluate the data sets for patterns and reduce information to a convenient form.

(ii) Inferential Statistics:

In Inferential statistics we use the sample data to make estimates or predictions about a large set of data (also known as population or universe) and test their suitability.

1.4 Universe/Population

The aggregate collection or whole group of individuals or objects possessing certain common characteristics which is of the interest of study is called a population or universe, e.g., population of some college students, population of library books, population of biscuit factories, etc. Sample is only a part or fraction of population. It is selected with an object of drawing inferences about the various population characteristics (parameters). Each population unit open for sampling is termed as sampling unit and of them units selected in the sample are called sample units.

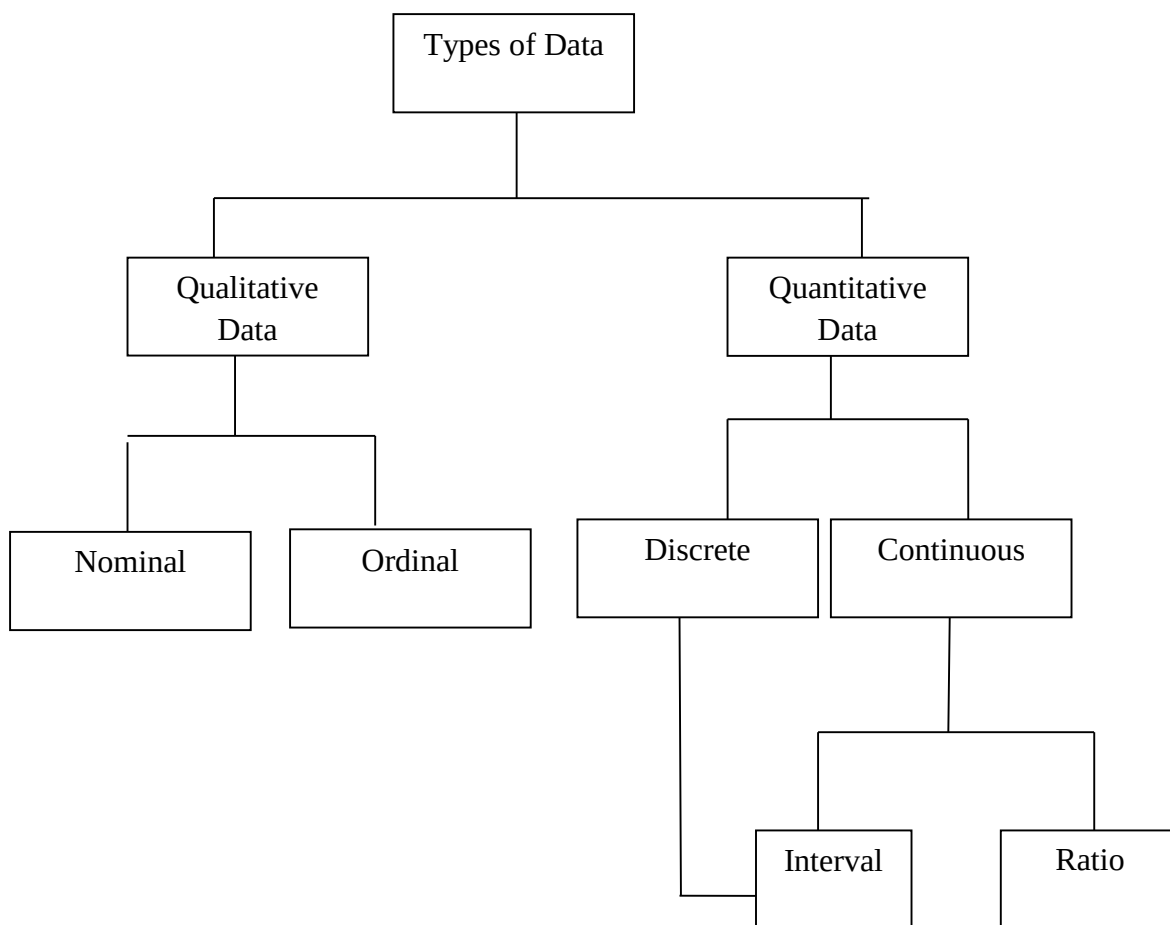
1.5 Statistical Problems/Limitations

Despite of vast use of statistics in different dimensions, there are certain limitations of the Statistics and Statistical Methods, which we may call as Statistical problems. These stated as follows-

1. Statistical laws are not exact laws like mathematical or chemical laws. They are derived by taking a majority of cases and are not true for every individual. Thus statistical inferences are uncertain.
2. Statistical methods deal with population or aggregate of individuals rather than individuals. When we say the average height of an Indian is 160 cms, it does not show height of an individual but as found by the study of average or an aggregate of individuals.

3. Statistical techniques apply generally to data which are reducible to quantitative forms. Consequently, the characteristics which cannot be expressed in figures cannot be studied satisfactorily. Such characteristics are beauty, goodness, health, intelligence, honesty, etc.
4. Statistical results might lead to fallacious conclusion if they are quoted short of their context. The argument that “in a country 15000 vaccinated persons died of small pox. Therefore, vaccination is useless” is statistically defective, since we are not told what percentage of the persons who were not vaccinated and had died.
5. Statistical technique is the same for the social and physical sciences, while both are different in nature.
6. Only one, who has as expert knowledge of Statistical methods can handle the Statistical data property. The data placed in the hands of an inexpert may lead to fallacious results or wrong conclusions.

1.6 Measurement and Scales



Characteristics under study termed as “variable” is a quantitative or numerically expressed qualitative characteristic which varies from one object to another within its domain. Any variable of interest is measured on the units under study to generate the observations known as **Statistical data**. Hence we may say that Statistical data refer to the numerical description or measurement of quantitative aspects of things under observations. For example, number of students in a class, number of colleges in a city, temperature, rainfall, etc. Observations or statistical data may be measured according to their classification shown in a diagram below:

1.7 Measurement of Qualitative Data:

For a quantitative characteristic called attribute we can simply observe or note their presence or absence in under observation. There is no natural numeric scale for its measurement. For example, gender, eye, color, beauty, etc. cannot be measured in number of numerically. In such situations we use two types of measurements known as nominal and ordinal scales and resulting into nominal and ordinal data according to their type of measurements.

Nominal Scale

To classify characteristic of people, objects or events into categories under some name is known as nominal scale. For example, gender is classified under name of the male and female color may be classified as black and white, etc.

Ordinal Scale (Ranking Scale):

Characteristic which can be put under order categories measured on ordinal scale are known as ranking scale and data thus generated are known as Ordinal Data. For example, socio-economic status may be measured as low, medium or high status.

Qualitative Data:

Statistical data which refers to the numerical description of the character under study of things under observations is known as quantitative data. This description may be in the form of counts or measurements. For example, number of students in a class, and separate counts for various kinds such as male and female students. These counts refer to discrete types of variable. The observation

may also include measurements as heights and weights, which are referred as continuous variable. The type of variable classifies the type of data as well. There is natural numeric scale to measure discrete variable or continuous variable such as age, height, weight which is expressible in numbers. It can be measured according to two scale known as interval and ratio scale of measurement.

Interval Scale

The interval scale is more sophisticated than nominal or ordinal scale. This scale can not only be ordered but the distance between two measurements can be obtained. The distance between these ordered category values are equal because there is some acceptable physical unit of measurement. However, the zero point is arbitrary. It can take continuous or discrete values.

Example: Fahrenheit (or Celsius) scale for measuring temperature where a temperature of 0°F , does not mean there is no heat. In fact, there is still some heat at temperature 10°F , -20°F , and so on. Because there is still some heat the (variable being measured) when zero is assigned as the measurement, the zero is not absolute zero.

Ratio Scale:

The highest level of measurement is the ratio scale. This scale has a true zero point and has the “equal ratio” properties. It consists of meaningful ordered characteristics with equal intervals between them. Presence of zero point is not arbitrary and is absolute. It is possible to multiply or divide across a ratio scale. Ratio between two values on the scale is a meaningful measure of the relative magnitude of the two measurements. The reason for the name ratio makes sense to say that a line that is 2.5 cms long is half the length of a 5 cms line. Similarly, it makes sense to say that a 20 second is twice the duration of 10 seconds.

Check your progress

1. Explain the population.
2. What do you mean by measurement of qualitative data?

1.8 Methods of Data Collection

Once it is decided what type of study is to be made, it becomes necessary to collect information about the concerned study, mostly in the form of data. For this information has to be collected from certain individuals directly or indirectly. Such a technique is known as survey method. These are commonly used in social sciences, i.e. the problems relating to sociology, political sciences, psychology and various economic studies. In surveys the required information is supplied by the individual under study or is based on measurements of certain units.

Types of Data

There are two categories of data namely primary data and secondary data depending upon the method of its collection.

Primary data and its collection

The data which are collected from the units or individual respondents directly for the purpose of certain study or information are known as primary data.

Secondary Data and its collection

It is the data which has been collected by certain people or agencies and statistically treated. Now the information contained in it is used again from the records processed and statistically analyzed to extract some information for other purpose. Usually secondary data is obtained from year book, census reports, survey reports, official records and reported experimental finding large scale data cannot be collected repeatedly because of the paucity of time, money and personal. Hence the use of secondary data for certain studies is inevitable. While making use of secondary data, one should always take care of the following points:

- (a) One should see whether data are suitable for study
- (b) The sources of data should also be viewed, keeping in mind whether at any time, it is reliable or not. If there is any doubt about the reliability of data, it should not be used.
- (c) It should be noted that the data are not obsolete.

- (d) In case the data based on a sample, one should see whether the sample is proper representative of the population.

Discrete and Continuous Data

Statistical data may be looked upon as a collection of facts, observation or information in numerical terms on variables under study regarding population/ universe or a sample from the universe to achieve the objectives of study or research. That variable which is capable of assuming every possible fractional value within its possible limits (called domain), when measured on different units is called continuous one; e.g. individual weight, height, age, rod-length, etc, Therefore, continuous data are those which have uninterrupted range of values and can assume either integral or fractional values.

A variable assuming certain specific or the integral values only when measured, is called the discrete or discontinuous one e.g. the number of members in a family, number of petals in a flower, number of fruits in a baskets and the like, So the discrete data are distinct, separate and invariable whole numbers, Statistical data are also called discrete or continuous data according to the nature of the variable they are associated with.

The statistical methods are applicable only when some data are available. The data can be quantitative as well as qualitative. If the data are qualitative they are quantified by using techniques like ranking, scoring, scaling or coding, etc. The data are collected either by experiment or by survey methods (directly or indirectly) and they are tabulated and analyzed statistically. Whatever may be the resulting value obtained from analysis proper and correct inferences have to be drawn from these numerical values. These inferences lead to final decision.

Preparation of Tables

Tabulation should not be confused with classification as the two differ in many ways, Mainly the purpose of classification is to divide the data into homogeneous groups or classes whereas the data are presented into rows and columns in tabulation. Hence classification is a preliminary step prior to tabulation. The following steps for the preparation of table are as follows:

1. The shape and size of the table contain the required number of rows and columns with stub and captions and the whole data should be accommodated within the cells formed corresponding to these rows and columns.
2. If a quantity is zero of the table should be entered as zero, leaving blank space or putting dash in place of zero is confusing and undesirable.
3. In case two or more figures are the same, ditto marks should not be used in a table in the place or original numerals.
4. The unit of measurement should either be given in parenthesis just below the columns captions or parenthesis along with the stubs in the row.
5. If any figure in a table has to specified for a particular purpose, it should be marked with an asterisk or a dagger. The specification of the marked figure should be explained at the foot of the table with the same mark.

Processing Classification of Data

Before tabulation of primary data, it should be edited for (i) consistency (ii) accuracy and (iii) homogeneity

Consistency:

Some information given by the respondent may not be compatible in the sense that information furnished by the individual either does not justify some other information or is contradictory to earlier one, For example the total expenditure exceeds the total income reported by the respondent the number of children mentioned is less than total number of sons and daughters, and then respondent should again be contacted to rectify the mistake so that there may be consistency in data.

Accuracy:

Accuracy is of vital importance. If the data are inaccurate, the conclusions drawn from it have no relevance or reliability. By checking the schedules and questionnaires only a little improvement can be made. For example, if the sum of certain figure is wrong it can be corrected but if the investigation has either made a false report or the respondent has deliberately supplied wrong

information about his income, age assets etc. editing will be of no use. In recent times, checks have been involved to attain accuracy e.g. by sending supervisors to check work of investigators or reinvestigating a few respondents after a certain gap of time.

Homogeneity:

To maintain the homogeneity, the information sheets are checked to see whether the unit of information or measurements is the same in all the schedules. For instance, some people might have reported income per month and some annual income. In such a situation it has to be converted to the same unit during editing. It should also be checked whether or not the same information has been supplied for a particular question in all the information sheets. The ambiguity arises due to various interpretations of same questions and should be removed. Once the primary data have undergone the above process it is fit for further analysis.

1.9 Summary

The word “Statistics” meaning “status” and “an organized political state” is derived from German or Latin. Statistics as a discipline may be defined as the science of collection, presentation, analysis and interpretation of numerical data and it is applied in various fields like business, commerce, industry, government, biological sciences, social science, agricultural sciences etc.

The aggregate collection or whole group of individuals or objects possessing certain common characteristics which is of the interest of study is called a population or universe. The characteristics under study are called variable. Limitation of Statics is that statistical results are applicable only on group and not on individuals. There are different measurement scales known as nominal, ordinal, interval and ratio scales.

Depending upon the finite or infinite number of values taken by variables they are classified as discrete or continuous. Data collected directly by the method of interview, measurement or questionnaire, etc is called primary data whereas those taken from previous records are called secondary data.

1.10 Self-Assessment Questions

1. Give different senses in which word “statistics” is used.
2. Describe the scope of statistical methods and specify their limitations.
3. Which of the following are statistical statements? Give reason.
 - Shakespeare was a great poet.
 - The average age of students of this Institute is 20 years.
 - The production of sugar in a certain district was quintals per acre in a particular year.
4. Comment on following
 - Statistics can prove anything.
 - Figures won't lie but lairs figure.
5. What is statistical data and how it is classified?
6. Differentiate between the following:
 - Quantitative and qualitative data
 - Discrete and continuous data
 - Nominal and ordinal scale
 - Interval and ratio scale
 - Primary and secondary data
7. Discuss various methods of measurement of data?
8. Write a note on methods of collection of data?
9. State briefly the advantages of “secondary data” over “primary data”.
10. What are the precautions to be taken while handling the secondary data?

1.11 Further Readings

1. Goon, A.M. Gupta M.K. & B. Dasgupta: Fundamentals of Statistics, Vol. 1, The World Press Pvt. Ltd., Calcutta.
2. Yule G.U. and Kandall, M.G.: An Introduction to the Theory of Statistics, Charles Griffin and Company Limited.
3. Weatherburn, C.E.: Mathematical Statistics.

Unit-2: Representation of Data-I (Diagrammatical Representation)

Structure

- 2.1 Introduction**
- 2.2 Objectives**
- 2.3 Frequency Distribution**
- 2.4 Diagrammatical Representation**
- 2.5 Bar Diagram**
- 2.6 Multiple Bar Diagram**
- 2.7 Sub-Divided Bar Diagram**
- 2.8 Percentage Bar Diagram**
- 2.9 Pie Chart**
- 2.10 Pictogram**
- 2.11 Leaf Chart**
- 2.12 Summary**
- 2.13 Self-Assessment Exercises**
- 2.14 Further Readings**

2.1 Introduction

We have seen in the last unit how Statistics and Statistical Methods provide a valuable assistance for the study, solution and formation of different kind of problems in almost all spheres of human activities. The statistical data are collected either by experiment or by survey methods (directly or indirectly). The way data is coming as and when observed, these data follow no order and are offered perhaps the way originally reported. There may be a complete lack of any systematic arrangement by size or sequence. Such unorganized data are known as *raw data* or *ungrouped data*. The data collected through surveys or experiments is the raw data and will be in a haphazard and unsystematic form. Such a data is not appropriately formed to draw right conclusions about the group or population under study. Hence it becomes necessary to arrange or organize data in a form, which is suitable for identifying the number of units belonging to a more classified group for comparison and for further statistical treatment or analysis of data.

Statistical data which usually refers to the numerical descriptions of the character under study of units or things under observations is known as quantitative data and may be in the form of counts or measurements. For example, number of members in a family, and separate counts for various kinds such as male and female. These counts refer to *discrete type* of data. The observation may also include measurements as heights and weights, which are referred as *continuous data*. There is a natural numeric scale to measure discrete variable or continuous variable such as age, height, weight which is expressible in numbers. The placement of these data in different homogeneous groups, formed on the basis of some characteristics or criterion is called classification or tabulation of data leading to better understanding and statistical analysis.

Let us consider an example where the raw data is collected about the group or population under study. For instance, the people may be divided into different age groups like <10, 10-20, 20-30, 30-40, etc. Or may be classified according to their monthly income (in rupees) like <500, 500-1000, 1001-2000, etc. Further these classified data can be presented in the form of well-arranged tables. These tables depict clearly the values or number of units possessing the required characters or belonging to specified classes. Another type of classification is the time series data in which data or the derived value form data for each time period is arranged chronologically.

2.2 Objective:

After going through this unit you should be able to know about-

- Frequency Distribution
- Diagrammatical Representation
- Pie Chart
- Bar Diagram
- Divided Bar Diagram
- Percentage Bar Diagram
- Pictogram
- Leaf Chart

2.3 Frequency Distribution

The premiere of data in form of frequency distribution describes the basic pattern which the data assumes in the mass. Frequency distribution gives a better picture of patterns of data if the number of items is large enough.

From a frequency array, it is not possible to compare characteristics of different groups. Hence for this, the classes are established to make the series of data more compact and understandable. The width of a class i.e. the difference between the upper and the lower limit of the class is termed interval. Once the classes are formed, the frequencies for these classes from raw data are expedited with the help of little slanting vertical strokes called tally marks. A bunch of four tally marks is crossed by the fifth to make the counting simpler. The whole process is as follows-

Defining the task

Let us assume that we have a set of data collected through a sample survey, which consists of a given number of observations on a certain quantitative variable. They differ in magnitude but in the manner presented these raw observations do not exhibit any sensible pattern and therefore the first and foremost task in dealing with such data is to arrange them in right format. The format should aim at

- (i) Organizing data in a manner that these become easy to read, understand and assimilate, and
- (ii) Summarizing data in a way that the basic trends and broad variations come to the fore and get highlighted.

When presented in the resultant form a researcher or an analyst gets a better grasp of the data. 'It facilitates a more efficient data analysis, which helps quicken the process of decision making.

Example

Consider the following raw data given in table which refers to weekly earnings of 80 female workers engaged in weaving trade at Surat during a particular year.

Weekly earnings of 80 female workers engaged in weaving trade at Surat

1052	1088	1077	1078	1089	1089	1082	1084	1088	1090
1099	1101	1102	1055	1063	1073	1078	1113	1086	1089
1080	1095	1092	1103	1118	1098	1097	1081	1061	1080
1083	1079	1111	1064	1056	1068	1055	1073	1075	1083
1085	1086	1083	1090	1105	1090	1069	1058	1072	1073
1086	1071	1070	1065	1059	1080	1084	1085	1075	1064
1087	1091	1108	1094	1097	1093	1107	1094	1082	1116
1085	1070	1076	1069	1061	1114	1089	1074	1105	1082

Preparation of Frequency Distribution (Tally Method)

To take a frequency distribution table for above example by using tally marks we proceed as follows-

1. Obtain the range of the distribution as the difference between the lowest and highest observation(s) For the data listed in table 1052 is the lowest and 1118 is the highest observation, with $(1118-1052=66)$ as the range of the distribution.
2. The range is then divided into an appropriate number C, which represent the width of the class interval. This also determines the number of class interval k among which individual

observations are distributed. If $C=10$, the range 66 is divided by 10 resulting in 6.6 n rounding to the next higher digit, it gives $k=7$ class intervals.

3. After completing step 2), all individual observations in the original data are picked up one by one and tally bar is marked opposite the class in which a particular observation falls. For example, in table an observation 1052 lies in the class (1052-1059) so that a tally bar is marked against this class. This has to go on till all the observations have been recorded by making tally marks.
4. Finally tallies marked against each class are counted and their total number recorded under a separate column heading frequency, as is column (3). For convenience in counting the number of tallies entered in each class, every fifty tally mark crosses the earlier our tallies diagonally from top to the bottom. Adding all class frequency yield a number 80 equal to the total number of observation (N) so recorded.
5. Frequencies obtained in column 3 may be expressed as present class frequencies as shown in column 4.

Frequency distribution of weekly earning of 80 female workers engaged in weaving trade at Surat (Tally Method)

Class-limits 'C' (1)	Tally Marks (2)	Frequencies 'f' (3)	Percentage class frequencies (4)
1050-1059		6	7.50
1060-1069		9	11.25
1070-1079		15	18.75
1080-1089		25	31.25
1090-1099		13	16.25
1100-1109		7	8.75
1110-1119		5	6.25
Total		80	100

The distribution constituted by column 1 and 3 in the above table is known as frequency distribution. It gives the number of women according to their weekly earnings. For example, 6 women's earning is between Rs 1050 to Rs. 1059; 9 women's earning is between Rs. 1060 to Rs. 1069 and so on. This frequency distribution has helped to understand and analyze the haphazard data in a systematic manner which is easy to handle for further treatment.

Smoothing of a grouped Distribution (Inclusive and Exclusive type Class Intervals)

When the upper limit of the previous class is not as the lower limit of the following class, as in the above example, and both the class limits are included in the same class are called inclusive type class intervals. In such case the classes do not constitute the continuous distribution and has to be made continuous. The simplest way to do this is to find the difference of the upper limit of the preceding class and lower limit of the succeeding class. Subtract half of the difference from the lower limit of each class and add the same from its upper limit. Continue this process for all classes. In the resulting class intervals the upper limit of the previous class is same as the lower limit of the following class, and only lower class limit is included in the corresponding class not both the limits as in case of inclusive type of class intervals and are therefore called exclusive type class intervals.

Frequency distribution (Inclusive type class intervals) of the Weekly earning of women

Weekly earning (Class-limits)	Number of women (Frequencies)
1050-1059	6
1060-1069	9
1070-1079	15
1080-1089	25
1090-1099	13
1100-1109	7
1110-1119	5
Total	80

The given distribution is not continuous as the upper limit of the preceding class is not the same as lower limit of following class. Hence it is smoothened. The difference between 59 and 60 is 1. Therefore, 0.5 is to be subtracted from the lower limit of the classes and 0.5 is to be added to the upper limit of the classes. Since the difference is constant the same quantity is subtracted and added in all the classes.

Frequency distribution (Exclusive type class intervals) of the Weekly earning of women

Weekly earning (Class-limits)	Number of women (Frequencies)
1049.5-1059.5	6
1059.5-1069.5	9
1069.5-1079.5	15
1079.5-1089.5	25
1089.5-1099.5	13
1199.5-1109.5	7
1109.5-1119.5	5
Total	80

Open End Classes

An open end class is a class taking one limit. Generally it is the lowest class taking the lower limit and highest class taking the upper limit. For instance, in an age group distribution, the lowest class is taken as less than five (<5) and highest class as more than the seventy (>70). Open end classes make it possible to accommodate values which are at large gaps without increasing the number of consecutive classes. However, open end classes should be avoided as far as possible. Open ends create problem in processes like computation and graphical representations.

Cumulative Frequency

The frequencies may be added up or cumulated on either from top to bottom (on the less than basis) or from bottom to the top (on more than basis). Cumulative frequencies less than type are obtained by adding successive frequencies from top to bottom as given in col. (4). Those of 'more

than type' cumulative frequencies are obtained by adding successive frequencies form bottom to top as given in col. (5) as shown in the following table. The less than type cumulative frequencies correspond to the upper limit of the class whereas more than type cumulative frequencies correspond to the lower limit of the class.

Cumulative Frequency

Weekly earning (Class-limits)	Number of women (Frequencies)	Cumulative Frequencies (Less than type)	Cumulative Frequencies (More than type)
1049.5-1059.5	6	6	80
1059.5-1069.5	9	15	74
1069.5-1079.5	15	30	65
1079.5-1089.5	25	55	50
1089.5-1099.5	13	68	25
1199.5-1109.5	7	75	12
1109.5-1119.5	5	80	5
Total	80		

In the above example cumulative frequencies help in finding out total number of women whose earning is less than Rs. 1059.5 is given by 6 whereas total number of women whose earning is more than Rs. 1049.5 is given by 80, similarly total number of women whose earning is less than Rs. 1069.5 is given by 15 whereas total number of women whose earning is more than Rs. 1059.5 is given by 7.4 and so on.

2.4 Diagrammatical Representation

How so ever informative and well designed a table may be, the pictorial representation of data is definitely a better tool for conveying details of the data to the common man in a simpler and more

understandable manner. The figures given in tabular form as such are not easily intelligible because of their dull, confusing and dispelling nature. If they are large in number or in size, then their study needs much time and brings a strain upon the mind. Diagrams on the other hand, are attractive and catch the attention of the reader by explaining and exposing the significant facts given in the data in visual and summarized form. They have a more lasting effect on the brain. When data of two items are compared with one another it is always easier through diagrams and graphs. It is for this reason that the government, various business houses and institutions are producing popular versions of their important statistics these days in the form of multi-colored booklets full of pictures, geometrical figures, curves and maps, etc. So it is very useful to represent statistical data by means of a diagrams and graphs which make of the unwieldy data intelligible and convey to the eye the important characteristics, general tendency and trend of the observations. It is now an essential part of the analysis and presentation process of statistical data. The four basic purposes of diagrams/graphs are to

- (1) Compare “proportions & relative changes” in data
- (2) Show trends/tendencies of data
- (3) Study how response changes over time in given set of data
- (4) Indicate how one variable relates with one another.

For a graph, a proper title, labeling for the axes and units of measurement are important. A good graph is designed so that it gives brief description at a glance. Similar descriptions should be provided for diagrams also.

However, there are certain disadvantages also. Graphs do not give accurate measurement of the variables as are given by tables. The numerical value can be obtained to any number of decimal places but from graph it cannot be found say to 2nd or 3rd places of decimals. Another disadvantage is that it is very difficult to have a proper selection of scale. By different scales it is possible that the facts may be misrepresented.

Different Categories of Diagrams

In this section we are going to deal with the following diagrams and graph to represent different types of statistical data.

- Bar Diagram
- Divided Bar Diagram
- Percentage Bar Diagram
- Pie Chart
- Pictogram
- Leaf Chart.

2.5 Bar Diagram

As the name indicates the bar diagrams are the simplest one dimensional diagram. A bar diagram is a visual display used to compare the frequency of occurrence of different characteristics of categorical data. They are particularly used if items of the different classes are not component parts' of the whole but are related with each other by their possession of some common characteristics. Bar height commonly represents a count of cases or frequencies for each category, a percentage of the total number of cases, or a function of another variable (e.g. the mean value for each category). Here thickness of the bar had to do nothing with the interpretation of the figures for which only the length or height is taken into account.

Bars are spaced at an equal interval. As far as possible they should be placed in ascending or in descending order of their lengths, allowing half bar width between bars and at each end vertical scale should start only with "0" only.

The scale should be mentioned on the diagram but it should be quite a convenient scale. The main purpose of the bar graph is to

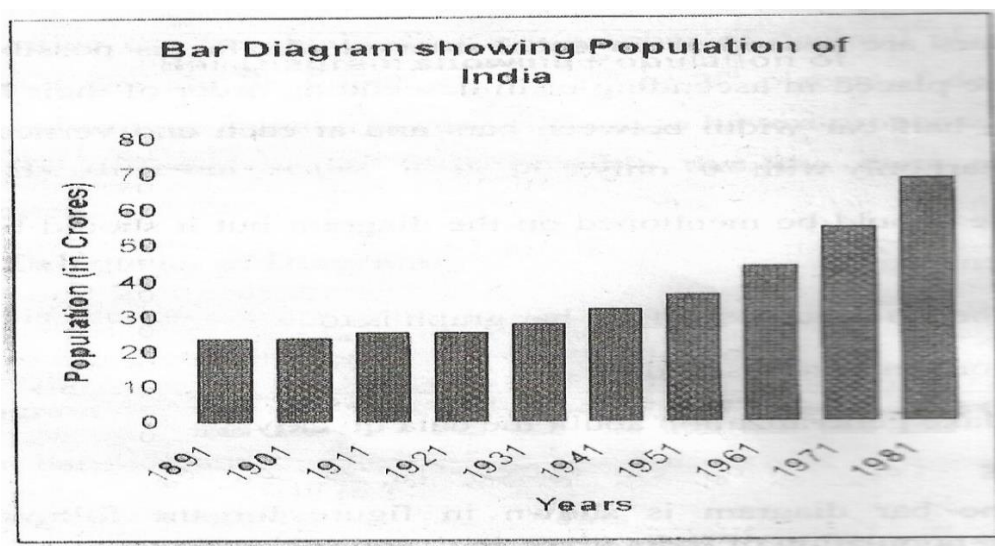
1. Compare groups of data
2. Make generalization about the data quickly.

Example: The bar diagram is shown in figure for the following data regarding population of India given in the table.

Table

Year of Census	Population (in crores)
1891	23.59
1901	23.83
1911	25.20
1921	25.13
1931	27.89
1941	31.86
1951	36.10
1961	43.91
1971	54.82
1981	68.38

Showing Bar data given in



Check your progress

1. Explain the frequency distribution.
2. What do you mean by Bar diagram?

2.6 Multiple Bar Diagram

They are extended form of the simple bar diagram. Here more than one aspect of the data is presented simultaneously. Each aspect is shown with different shades or colors. The bars of one groups are separated from other groups by putting them adjacent. Also the multiple bars can be shown as placed on one another without loss of clarity to save space.

These diagrams are very useful for comparison between two or more phenomena by representing them with different bars having different shades or colors. An index explaining shades/colors and scales used should be shown in the diagram. Bars may be horizontal or vertical. The space between bars representing the components of the same total is taken smaller than space between bars of the different sets of total. However, the totals in themselves are not easily comparable here as though the simple bar diagrams. These diagrams are meant for comparing two or more sets of interrelated points of time, place or categories, etc.

Example

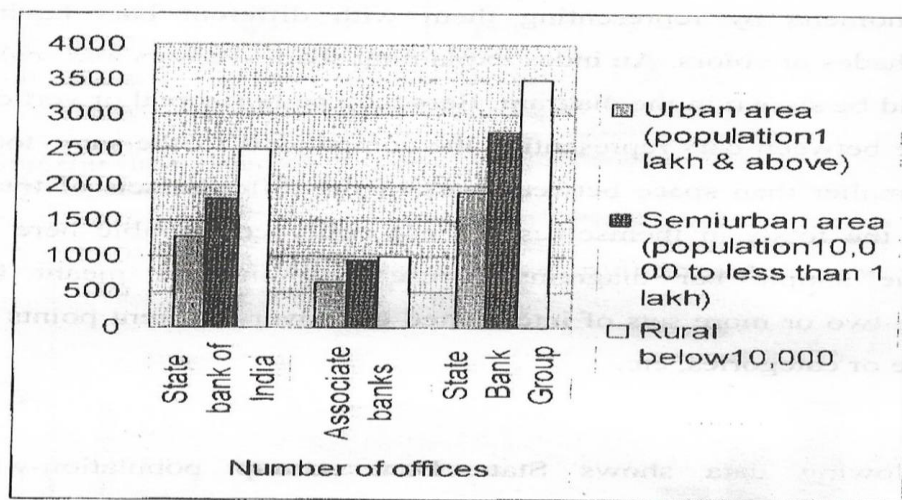
Following data shows State Bank Group population wise distribution of India's offices at end of 1980). Prepare a multiple Bar Diagram.

Table-Population wise distribution of Indian Offices of SBI

Population Category	Number of Offices		
	State Bank of India	Associate Banks	State Bank Group
Urban area (population 1 lakh & above	1270	632	1902

Semi urban area (Population 10,000 to less than 1 lakh)	1807	940	2747
Rural below 10,000	2492	999	3491

Figure showing Multiple Bar Diagram for above table



2.7 Subdivided Bar Diagram

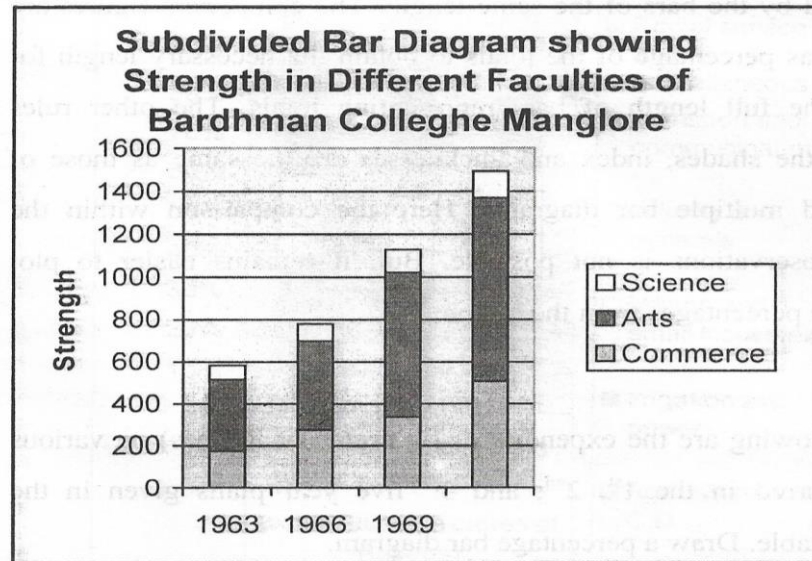
When it is desired to show the aggregates and their divisions into various components, the bars are drawn proportional in length to the totals and are subdivided into ratios of their components. Each subdivided part of the bar will correspond in size to the value of the item, it represents. Such diagrams are called Subdivided Bar Diagram. While preparing these diagrams, it must be observed that the arrangements of the various components remain identical for all the bars to avoid confusion and keep the diagram readily distinguishable. As usual, different shades or colors are to be used for representing different components of the total but shades of each component will remain the same for all the bars. Index of shades used should be shown with the diagram.

Example: Draw a subdivided bar diagram for the following data given in table.

Table

	Strength of students of Bardhman College Manglor in the year									
Faculties	1936	1964	1965	1966	1967	1968	1969	1970	1971	1972
Science	177	185	260	270	296	341	338	398	459	507
Arts	340	414	417	426	470	681	684	696	779	862
Commerce	65	85	82	84	84	109	106	87	103	125
Total	582	684	795	780	850	1131	1128	1181	1361	1494

Figure showing Sub-divided Bar Diagram for table



Thus we see from figure that these subdivided bar diagrams denote not only the variation in the total of the values of given characteristics but changes in the components parts of the total are also exhibited.

Limitations

Since the components of the bars do not start from the same scale value, so they do not remain easily comparable in their size across-sections. Here individual bars are to be studied separately and properly for the inter component comparisons.

2.8 Percentage Subdivided Bar Diagram/ Percentage Bar Diagram

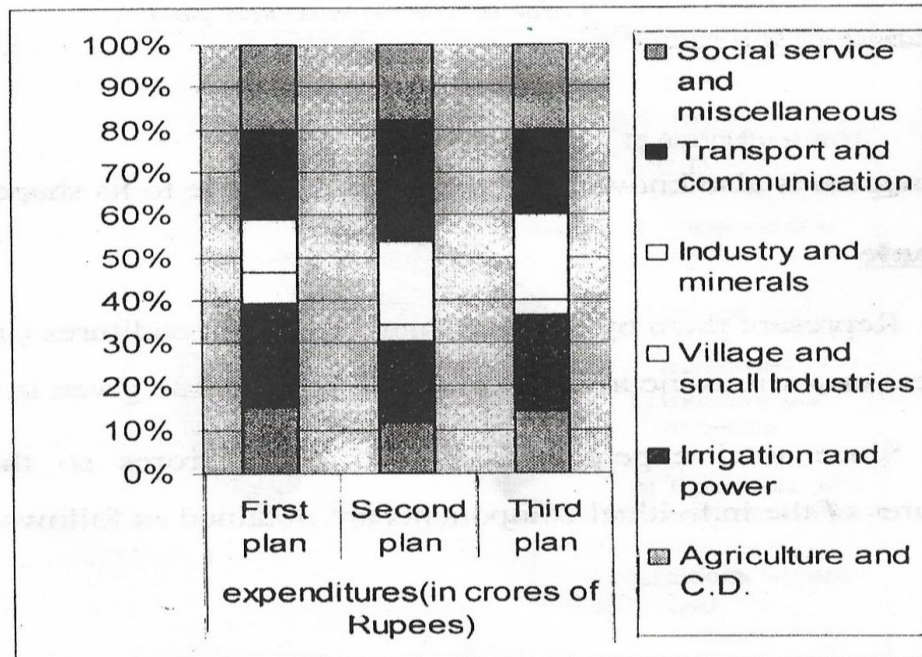
If the purpose of the graph is only to show the proportionate composition of the totals with respect to their component parts, it is best served by Percentage Subdivided Bar Diagram or simply by a Percentage Bar Diagram. Here all the totals are equated to 100% and are represented by the bars of the same length. The component figures are expressed as percentage of the totals to obtain the necessary length for them in the full length of bars representing totals. The other rules regarding the shades, index and thicknesses are the same as those of simple and multiple bar diagrams. Here the comparison within the original observations is not possible. But it remains easier to plot cumulative percentages from the bottom.

Example: Following are the expenditures (in crores of Rupees) on various heads incurred in the 1st, 2nd, and 3rd five year plans given in the following table. Draw a percentage bar diagram.

Table

Subject	Expenditures (in crores of Rupees)		
	First Plan	Second Plan	Third Plan
Agriculture and C.D.	361	529	1068
Irrigation and power	561	865	1662
Village and small industries	173	176	264
Industry and minerals	292	900	1520
Transport and communication	497	1300	1486
Social service and miscellaneous	477	830	1500
Total	2361	4600	7500

Figure showing Sub-divided Bar Diagram for table



2.9 Pie Chart

Circles are preferred than rectangles bars when the difference between totals to be compared is larger. Circles are drawn with their radii in proportional to the square roots of the values they represent. All the centers of the circles must lie in a straight line. If totals are the sum of the various components, then each circle may be divided into as many segments as are the component in its corresponding total. The area of the segment has the same percentage to the total area of the circle as the represented value has with its total figure. We know that the sum of the angles round the centre of a circle is 360 degrees.

Pie graph displays percentages. The circle of a pie graph represents 100%. Each portion that takes up space within the circle stands for a part of that 100%.

$$\text{The angle of the sector} = \frac{\text{Value of the represented part}}{\text{The whole quantity}} \times 360^\circ$$

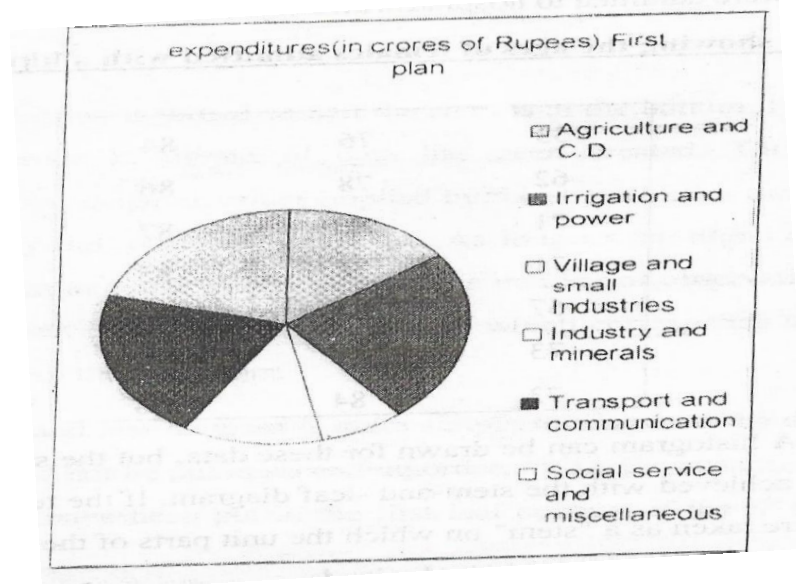
Pie diagram is also known as Circular diagram due to its shape.

Example: Represent them by a pie diagram the expenditures (in crores of Rupees) on various incurred in the first five-year plan given in table. Since total expenditures is Rs. 2361 Crores so the angular measures of the individual components are obtained as follows-

Table

Subject	Expenditures (in crores of Rupees)	
	First Plan	Degrees
Agriculture and C.D.	361	55
Irrigation and power	561	86
Village and small Industries	173	26
Industry and minerals	292	45
Transport and communication	497	76
Social service and miscellaneous	477	72
Total	2361	360

Figure showing Pie chart/diagram for adjacent table



2.10 Pictograph

The device of pictures is being profusely used now for comparing statistical data. The pictorial representation of facts is very often used in various exhibitions, propaganda posters, etc. They present dull masses of figures in an interesting and attractive manner through the objects of daily picture. The image of the entire data is fixed in the mind of the observer by a mere glance at the picture. Relationship between figures and their comparison can be studied through pictograms much more easily than by studying huge mass of numerical data.

Check your progress

1. Explain the pie chart.
2. What do you mean by Pictograph?

2.11 Leaf chart

Drawing a histogram, discussed in unit III, can often be quite tedious and an alternative method can be employed when the raw data (original non-grouped) are available. Table shows the ages at which 21 females were admitted to hospital with a hip fracture. A histogram can be drawn for these data, but the same end result can be achieved with the stem and leaf diagram. If the tens part of the age were taken as a “stem” on which the unit parts of the age were to be attached like leaf, would be obtained.

Table showing the ages of female admitted with a hip fracture.

53	76	84
62	78	86
71	82	87
71	78	85
67	84	87
73	84	94
73	84	98

The stem is written down in the first column and usually a vertical line is drawn to separate the stem from the leaves. Here the leaf is the unit's part of the age and the leaves belonging to each stem are put in the next columns. Thus there is only one patient in her 50s at the age 53 and digit "3" goes in the first leaf column on to the stem "5" there are two age (62 and 67) that belong to the second stem of "6", and the units '2' and '7' go in the first and second leaf columns. In this way the leaves are added to the stems to give the complete diagram.

Stem and leaf diagram of the ages (years) of 21 female's patients with a hip fracture.

[illegible]

If the diagram is turned so that stem is at bottom it can be seen that a crude histogram of data has been created. The class corresponds to the range of values implied by the stem -in this example, from 50 to 59 years, 60 to 69 years, etc. As long as the digits of the leaves are written in equal width columns, the number of observation in each class is given by the length of the row, which corresponds to the height of the bar of the histogram.

The stem and leaf diagram is quite simple to draw and the digits for the leaves need not be put in ascending order. In fig. for instance, the '8' (for 98) could have been put in the first leaf column of the '9' stem and the '4' (for 94) could have been put in the second column. Thus once the appropriate stem- has been chosen, the can be filled in easily by just going through the data without even having to order them.

For a given set of data, choosing an appropriate stem is sometimes a process of trial and error, but there are a few tricks that can ease the task. Usually the number of stems is between five and twenty but this depends on having a sufficient number of values for each stem. Note that it is not possible to create three stems within a single tens digit, since the number of possible different leaves on each stem must be the same.

2.12 Summary

For a better presentation and efficient analysis statistical data are classified, summarized and tabulated in the form of Frequency distribution using tally marks. Data may be grouped in inclusive or exclusive type of class intervals. This frequency distribution can further be treated and cumulative frequency may be obtained. For better understanding these frequency distributions can be shown on graphs as histogram, frequency polygon, frequency curve and ogives. Statistical data are diagrammatically represented as chart, bar diagram, divided bar diagram, percentage bar diagram, pictogram and leaf chart for facilitating analysis and comparisons of data over person, place and time giving lasting and eye catching effects.

2.13 Self-Assessment Exercise

1. Mention the methods generally used in the collection of statistical data with precautions to be taken.
2. What are different parts of a table? How a frequency distribution table is prepared?
3. What do you mean by classification and tabulation of data?
4. What do you understand by diagrammatic representation of data? What are its main advantages and disadvantages?
5. Discuss various methods of graphical representations stating the situations where they may be optimally used. Also state their limitations precautions to be taken.
6. For the data given below, draw a) a simple bar chart for GDI as percent to GDP, and b) a single chart showing the share of public and private GDI as percent to GDP.

GDI at Current Market Prices (as percent to GDP)

	1983-84	1984-85	1985-86	1986-87	1987-88	1988-89
GDI	23.1	26.1	27.1	24.6	26.2	23.4
Public	8.2	8.8	7.6	7.0	6.7	6.6
Private	13.0	14.8	18.9	14.9	14.7	15.2

7. Using the data given below, draw a) component chart for absolute figures, b) component chart after obtaining percent data, and c) pie chart for percent data for 1975-76 and 1980-81.

Gross Saving at Current Prices (Rs. Crore)

Year	Household	Pvt. Sector	Public Sector	Total
1975-76	28058	5243	7642	40943
1976-77	41567	5125	7539	54231
1977-78	65519	5789	6981	78289
1978-79	78248	7983	7803	94034
1979-80	82145	11580	7562	101287
1980-81	100453	15642	5423	121518

2.14 Further Readings

1. Goon, A.M., Gupta, M.K., Dasgupta, B.: Fundamentals of Statistics Vol. 1, The World Press Pvt. Ltd., Calcutta.
2. Yule G.U. and Kerdall, M.G.: An Introduction to the Theory of Statistics, Charles Griffin and Company Limited.
3. Weatherburn, C.E.: Mathematical Statistics.

Unit-3: Representation of Data-2 (Graphical Representation)

Structure

3.1 Introduction

3.2 Objectives

3.3 Graphical Representation of Frequency Distribution

3.4 Histogram

3.5 Frequency Polygon

3.6 Frequency Curve

3.7 Ogives

3.8 Summary

3.9 Self-assessment questions

3.10 Further Readings

3.1 Introduction

In general, the statistical data are unwieldy and as such its various features cannot be understood clearly and readily, at a glance. It has to be reduced in a suitable form. The representation of the data in a tabular form by a frequency distribution is one of the techniques to achieve this objective. Normally the frequency distribution is not able to highlight various salient features of the data. A graphical representation of the frequency distribution is a powerful tool of data representation and interpretation. Hence, the data is represented by way of lines, curves, dots and bars etc. on a graph paper with variable values being put on the X-axis and frequencies on the Y-axis. The graphical representation of the data is an attractive and impressive way of representation that has a more lasting effect on the mind of the human beings than the tabular form representation of the data. The shape of the graph provides easy answer and ideas regarding the variation of data, skewness, peakedness at the top of the frequency curve, modes, extremes, outliers, spread of the data etc. inherent in the distribution of the data. Accordingly, frequency distribution graphs serve as an effective tool of a quick analysis and effective comparison between two or more distributions. The pattern of variations and the points of contrast become quite obvious when the graph of one distribution is superimposed over the other.

3.2 Objectives

After going through this unit you will be able to know and draw

- Histogram
- Frequency Polygon
- Frequency Curve
- Ogives

3.3 Graphical Representation of Frequency Distribution

Graphic Representation of Frequency Distribution is a powerful tool of data presentation and interpretation because the shape of graph provides easy answers to several important questions. Normally the frequency distribution, as a tabular representation is not able to highlight the essential characteristics of the data as apparently as its graphic presentation may do. The shape of graph offers an exact idea of the variations, its skewness, peakedness, modes, extremes, outliers, spread etc. inherent in the distribution of data. Accordingly, frequency distribution graphs serve as effective tools of a quick analysis and effective comparison between two or more distributions. The pattern of variations and the points of contrast become quite obvious when the graph of one frequency distribution is superimposed on the other. Following are important methods of graphical representation of Frequency Distributions.

3.4 Histogram

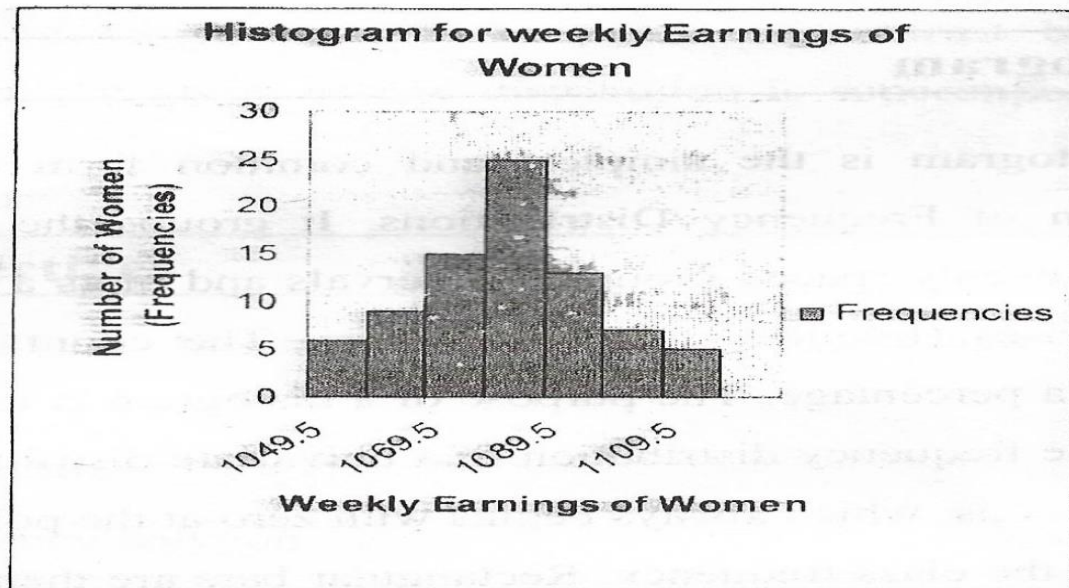
A histogram is the simplest and common form of graphical representation of frequency Distributions. It groups the values of a variable into evenly spaced or intervals and plots a count of the number of cases (frequency) in each group. The count can also be expressed as percentage. The purpose of a histogram is to graphically summarize the frequency distribution of a univariate distribution. Along the vertical Y-axis, which always begins with zero at the point of origin, are measured the class frequency. Rectangular bars are then raised over successive class intervals with their base equal in width on the X-axis. The height of each bar measured on Y-axis is kept equal to the corresponding class frequency. The area of the bar corresponding to each class interval is given by its class frequency “ r ” multiplied by the width of class interval C . Since frequency distributions may have equal or unequal class intervals, the procedure for drawing a histogram is described separately for both the situations are under.

(i) Histogram for Equal Intervals

Figure 3.1 represents histogram of the frequency distribution with equal interval given in Unit 2. The horizontal X-axis is divided by marking dots into equal parts numbering two or three more than the number of class intervals comprising the distribution. Starting from left not necessarily with zero, each dot is labeled by the lower class limit of the successive class, leaving a space equal

to the size of one class interval on the either extreme side. At times, the horizontal scale is also used to show the mid points of the successive class intervals.

Figure 3.1 showing Histogram for data given in table 1.2.4



(ii) Histogram for Unequal Class Intervals

Distribution with unequal class intervals is materially not different. It requires only minor adjustments in the spacing dots marked on the X-axis. Here frequency densities are plotted against class intervals instead of frequencies.

$$\text{Frequency density of any class} = \frac{\text{Class frequency}}{\text{width of the class}}$$

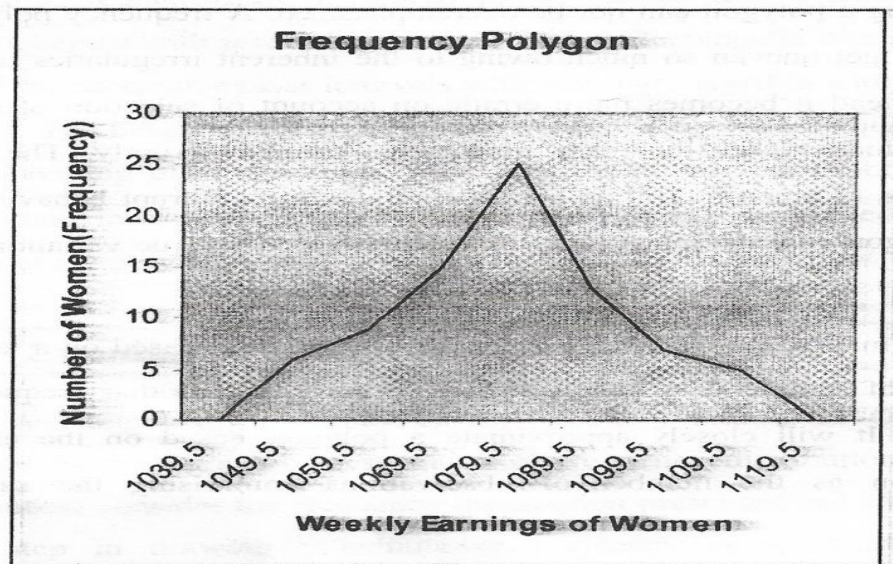
The method of calculation frequency density is simple and they are obtained by dividing the frequencies of each class by their respective class widths. A histogram for an open ended distribution is drawn essentially the same way, except that the open end class are not considered. Limits are chosen arbitrarily so that the width of the class interval becomes equal to the preceding (or succeeding) class.

3.5 Frequency Polygon

Frequency Polygon represents yet another way of depicting a frequency distribution in the form of a graph. Given the histogram in figure, a Frequency polygon is drawn by making dots at the mid-

points of the top of each bar joining them by means of straight lines. The polygon so obtained is closed at the end by joining the top base mid points of the first and the last rectangles with the mid points of the next outlying interval on either side. The mid-points of these two outlying intervals fall on their bottom base, meaning zero class frequencies.

Figure showing frequency polygon



In fact, constructing a frequency polygon does not necessarily require a histogram being drawn first. It can be obtained directly by plotting dots above each interval midpoint at heights equal to the corresponding class frequencies and joining them by means of straight lines. The polygon is closed on either side exactly the same way as explained above. However, the X-axis measures the successive class mid points, and not the lower class limits. This is shown in above figure.

Check your progress

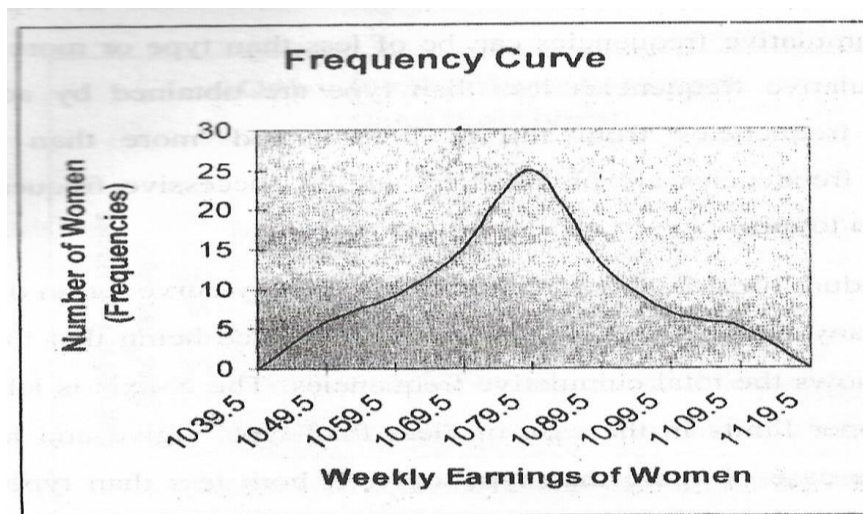
1. Explain the histogram.
2. What do you mean by frequency curve?
3. Define the frequency polygon.

3.6 Frequencies Curve

Frequency Curve are obtained by smoothing the frequency polygon drawing a free hand smooth curve through the various points that yield a frequency polygon on joining. Serious limitation of a smooth curve drawn in free hand is that no two persons will ever smooth the polygon in exactly the same way. Despite of this limitation, the need for smoothing a polygon cannot be overemphasized. A frequency polygon does not get uneven so much owing to the inherent irregularities in the data. Instead it becomes more erratic on account of selection of class width which makes the class frequencies change abruptly; the real advantage of smoothing thus lies in eliminating the abrupt behavior of the polygon and making it more representative of the true variations in the data.

It may be noticed that a frequency distribution based on a larger number of sample data observations will have a smoother frequency polygon, It will closely approximate a polygon based on the entire population as the number of observations comprising the sampe increases.

Figure showing frequency curve



A polygon may assume a variety of shapes, more frequently encountered among them are either symmetrical or skewed in shape. Some others not so common are J-shaped, V-shaped, S-shaped, bimodal, etc. Class frequency are measured along the vertical Y-axis, which always begins with zero at the point of origin. Rectangular bars are then raised for successive class intervals with their base equal in width on the X-axis. The height of each bar measured on Y-axis is kept equal to the

corresponding class frequency. The area of the bar corresponding to each class interval is given by its class frequency multiplied by the width of the class intervals.

3.7 Cumulative Frequency Curves or Ogives

A cumulative frequency curve, popularly known as Ogive, is another form of graphic presentation of frequency distribution. As an illustration consider the frequency distribution presented in Unit 2. The first step in drawing a cumulative frequency curve is to obtain cumulative frequencies denoted as ' cf ' and record them in separate column. Cumulative frequencies can be of less than type or more than type. Cumulative frequencies less than type are obtained by adding successive frequencies from top to bottom and more than type cumulative frequencies are obtained by adding successive frequencies from bottom to top.

Figure: Cumulative frequency curve or ogive (more than type)

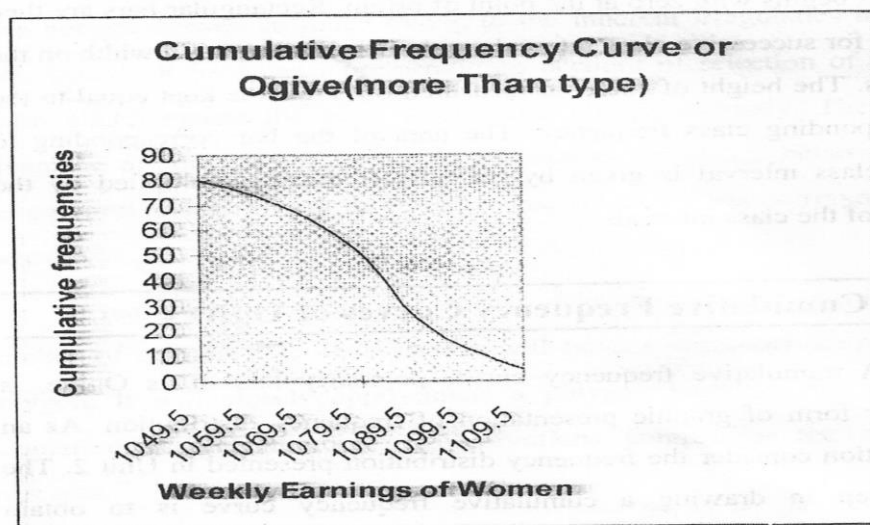
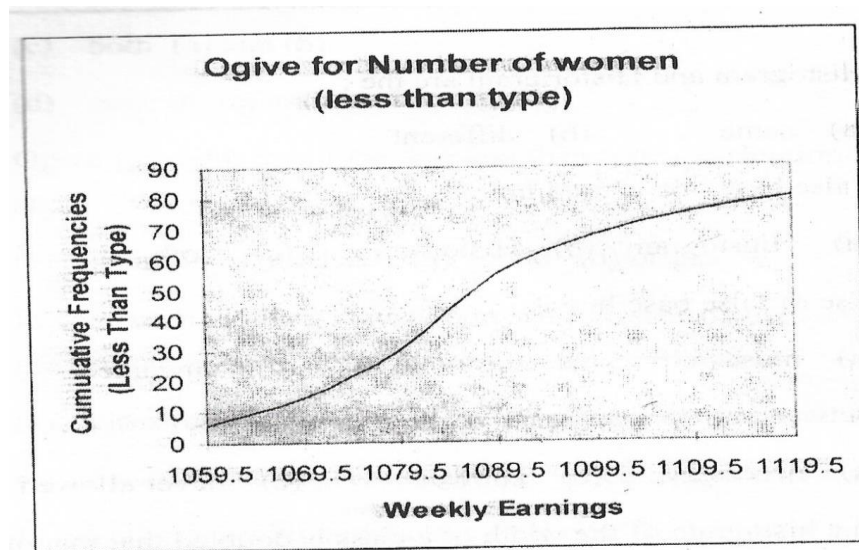


Figure: Cumulative frequency curve or ogive (less than type)

Procedure for drawing a cumulative frequency curve, or an ogive, is same as any frequency curve. The only difference being that the Y-axis now shows the total cumulative frequencies. The X-axis is labeled with the upper limits in the case of "less than type" ogive and lower limits in the case of 'more than type' ogive. If both less than type and more than type ogive are plotted on the same

graph they intersect at median of observations. The ogives can also be smoothed by free hand and frequency polygon.



3.8 Summary

Geographical Representation of frequency distribution is a powerful tool of data representation and interpretation because the shape of graph provides easy answer to several important questions. A histogram is the simplest and common form of graphical representation of data. Frequently polygon and frequency curves are other important tools of graphical representation. A polygon may assume a variety of shapes, more frequently encountered among them are either symmetrical or skewed in shape. Ogives are another form of graphic representation of frequency distribution. The point of intersection of two ogives gives the median. In a histogram the area of the rectangles equals the corresponding frequencies whereas in bar diagrams height of the bars equals the frequency.

3.9 Self-Assessment Exercises

- P-1 Histogram and Historigram are the: (a) Same (b) Different
- P-2 False base line is used in : (a) Histogram (b) Histoigram (c) both
- P-3 Use of false base line is:

- (a) Must (b) Desirable (c) unwanted
- P-4 Between two rectangles of a histogram a gap is:
- (a) necessary (b) allowed (c) never allowed
- P-5 In a histogram, if the width of a class is doubled that of other classes, then its frequency is:
- (a) doubled (b) halved (c) no change
- P-6 Class limits and class boundaries are:
- (a) Always same (b) Always different (c) not always different
- P-7 A time series data is presented by means of:
- (a) Histogram (b) Histogram (c) bar diagram (d) ogive
- P-8 The two types of ogives cut each other at:
- (a) Median (b) Q_1 (c) Q_3 (d) Mean
- P-9 A histogram is a:
- (a) diagram (b) graph (c) table (d) text
- P-10 Ogive curve occurs for:
- (a) More than type distribution (b) less than type distribution
- (c) both (a) and (b) (d) none of (a) and (b)
- P-11 Ogive for more than type and less than type distribution intersect at :
- (a) mean (b) median (c) mode (d) origin
- P-12 In case of frequency distribution with classes of unequal widths, the heights of bars of a histogram are proportional to:
- (a) class frequency (b) class intervals (c) frequencies in percentage (d) Frequency densities
- P-13 In a histogram with equal class intervals, the heights of rectangles are proportional to:
- (a) mid-values of the classes respective (b) frequencies of the classes
- (c) Either (a) or (b) (d) Neither (a) or (b).

3.10 Further readings

1. Goon , A.M., Gupta, M.K., Dasgupta, B. : Fundamentals of Statistics Vol. 1, The World Press Pvt. Ltd., Calcutta.
2. Yule G.U. and Kerdall, M.G.: An Introduction to the Theory of Statistics, Charles Griffin and Company Limited.
3. Weatherburn, C.E.: Mathematical Statistics.



**U. P. Rajarshi Tandon
Open University**

Master of Science PGMM -108N Mathematical Statistics

Block

2 Measures of Central Tendency and Dispersion

Unit- 4	Measures of Central Tendency-I	69-100
Unit- 5	Measures of Central Tendency-II	101-122
Unit- 6	Measures of Dispersion	123-148

Block-2

Measures of Central Tendency and Dispersion

The **Block - 2 – Measures of Central Tendency and Dispersion** is the second block with two units and deals with the different measures of central tendency and measures of dispersion. The *first unit* of this block deals with measures of central tendency. Once data have been collected and represented, one may like to know the particular value around which the data has the tendency to concentrate. This value is known as measure of central tendency. Various measures of central tendency along with their characteristics have been discussed in the *second unit*.

The *third unit* of this block deals with measures of dispersion. Once data have been represented and a measure of central tendency has been located, one may like to know the scatterness of the data around this measure of central tendency. Various measures of dispersion have been defined and their characteristics have also been discussed.

Unit-4: Measures of Central Tendency-I

Structure

4.1 Introduction

4.2 Objectives

4.3 Arithmetic Mean Short Cut Method

4.4.1 Grouped Data (Discrete Frequency Distribution)

4.4.2 Grouped Data (Continuous Frequency Distribution)

4.4.3 Properties of Arithmetic Mean

4.4.4 Properties and Advantages of Mean

4.4.5 Limitations of Mean

4.5 Geometric Mean

4.6 Harmonic Mean

4.7 Exercises

4.8 Summary

4.4 Further Readings

4.1 Introduction

Statistical Methodology is a comprehensive term which includes almost all the methods involves in the collection, processing condensing and analyzing of data. The data collected from the yield for a number of items vary greatly in their qualitative as well as quantitative nature. For example, the rainfall at a particular region is erratic in nature and shows variation from year to year, month to month and even day to day. The condensation of data in terms of maps, charts, diagrams, etc. is a first and necessary step in rendering a long series of observation comprehensible. But for practical purpose it is not enough, particularly when we want to compare two or more different series of data, e.g. we may wish to compare the distribution of status in two races of man, or the birth rates in India in two successive decades or rainfall in two different regions, or the number of wealthy people in two different countries. For such problems, there are certain statistical techniques one of which is a measure of central tendency.

In this unit have been highlighted different measures of central tendency are covered. Various situations where they find calculation of these measures for ungrouped and grouped data are described.

4.2 Objectives

After studying this unit, you will be able to-

- Understand the meaning of central tendency of data
- Understand the arithmetic mean short cut method
- Understand the geometric and harmonic mean.

4.3 Arithmetic Mean (Ungrouped Data)

The arithmetic mean of a series of n observations $X_1, X_2, X_3, \dots, X_n$ is obtained by summing up the values of all the observations and dividing the total by the number of observations. Thus,

$$X = \frac{\text{Sum of observations (or values)}}{\text{Number of observations}} = \frac{X_1 + X_2 + X_3 \dots \dots + X_n}{n} = \frac{\sum_{i=1}^n X_i}{n}$$

Where $\sum$ (sigma) stands for summations and X_i is the i th value of the observation (variable).

Example 1.1: The rainfall records in a month of 10 regions of a State is given below. Compute the average rainfall of the month for the State.

Region	1	2	3	4	5	6	7	8	9	10
Rainfall	17.6	10.1	11.4	18.5	10.5	14.3	8.9	13.4	10.6	12.5

(in mm)

Solution:
$$\begin{aligned} \text{Mean } (X) &= \frac{\sum_{i=1}^{10} X_i}{10} \\ &= \frac{X_1 + X_2 + X_3 \dots \dots + X_{10}}{10} \\ &= \frac{17.6+10.1+11.4+\dots+10.6+12.5}{10} \\ &= \frac{127.8}{10} = 12.78mm \end{aligned}$$

The computation of Arithmetic mean using short cut method is discussed below:

4.3.1 Short-Cut Method

This method is applied to avoid lengthy calculations. When the individual set of reading are large in size, an arbitrary value is selected as a working mean (known as assumed mean) and differences between the working mean and the individual readings (known as deviations from the assumed

mean) are worked out. By summing these differences dividing by the number of readings we can get the mean of deviations from assumed mean. Let $X_1 + X_2 + X_3 \dots \dots + X_n$ be n individual reading on the variable and let A be the working mean. Let $d_1, d_2, d_3, \dots, d_n$ denote the differences between the working mean and individual values $X_1 + X_2 + X_3 \dots \dots + X_n$ respectively. The mean $\bar{X}$ of X , in terms of the mean $\bar{d}$ of differences, is calculated as:

$$\begin{aligned} d &= \frac{d_1 + d_2 + d_3 \dots \dots + d_n}{n} \\ &= \frac{(X_1 - A) + (X_2 - A) + (X_3 - A) + \dots \dots + (X_n - A)}{n} \\ &= \frac{\sum_{i=1}^n X_i - nA}{n} = \frac{\sum_{i=1}^n X_i}{n} - A \\ &= \bar{X} - A \end{aligned}$$

or,
$$\bar{X} = A + \bar{d} = A + \frac{1}{n} \sum_{i=1}^n d_i$$

True mean guessed mean + (sum of deviations from guessed mean/number of cases)

This is useful, if the size of frequencies is large. An illustration is given below.

Example 1.2: Calculation the mean for the following scores: 60, 65, 74, 85, 95.

Solution:

Table 2.1 Distribution of Scores

X_i (Scores)	$X_i - 74$
60	-14
65	-9
74	0
85	+11
95	+21
	+9

4.3.2 Grouped data (Discrete Frequency Distribution)

In a discrete series, let the individual readings $X_1 + X_2 + X_3 \dots \dots + X_n$ of the variable X occur (have frequencies) $f_1 + f_2 + f_3 \dots \dots + f_n$ times respectively. The product X_i, f_i is the sum of all X_i 's in the data $\sum_{i=1}^n X_i, f_i$ and is the sum of all the observations. Then the mean of X is obtained by summing the product of individual readings with corresponding frequencies and dividing the total by the sum of frequencies, i.e.,

$$\begin{aligned} \text{Mean } \bar{X} &= \frac{X_1 + X_2 + X_3 \dots \dots + X_n}{f_1 + f_2 + f_3 \dots \dots + f_n} \\ &= \frac{\sum_{i=1}^n X_i, f_i}{\sum_{i=1}^n f_i} \text{ or } \frac{\sum_{i=1}^n X_i, f_i}{N} \end{aligned}$$

Where $N = \sum_{i=1}^n f_i$ is the total number of observations.

The various steps involved in the calculations of $\bar{X}$ under this method are:

- (i) Multiply the frequency of each row with the concerned variable (X, f) and total them.
It will be $(\sum Xf)$.
- (ii) Add up all the frequencies $(\sum f)$
- (iii) Divide $\sum Xf$ by $\sum f$ or N find out the arithmetic mean $(\bar{X})$

Short cut method:

To find our arithmetic mean in a discrete series by this method the following steps are taken:

- (i) Any value of the distribution may be taken as working mean or arbitrary mean, say A (preferably it should be near the middle of the frequency distribution).
- (ii) Take deviation d_x of the variable X from the working mean $(X-A)$ and denote it by d_x .
- (iii) Multiply each d_x by its respective f and denote it by fd_x .
- (iv) The arithmetic mean $(\bar{X})$ is then calculated with the help of following formula:

$$\text{mean} = \bar{X} = A + \frac{1}{n} \sum_{i=1}^n f_i d_{xi}$$

Check your progress

1. Explain the arithmetic mean.
2. What do you mean by discrete frequency distribution?

Example 1.3: Below are given the number of children born per family in 735 families in a locality. Calculate the average number of children born per family in the locality.

Table- 1.2
Number of Children Born per Family

Number of children born per family	Number of families
0	96
1	108
2	154
3	126
4	95
5	62
6	45
7	20
8	11
9	6
10	5
11	5
12	1
13	1

Solution: Computation of the average number of children born per family:

Table 1.3
Showing the calculation

Number of children born per family (X)	Number of families (f)	$X.f$ (xf)
0	96	$0 \times 96 = 0$
1	108	$1 \times 108 = 108$
2	154	$2 \times 154 = 308$
3	126	$3 \times 126 = 378$
4	95	$4 \times 95 = 380$
5	62	$5 \times 62 = 310$
6	45	$6 \times 45 = 270$
7	20	$7 \times 20 = 140$
8	11	$8 \times 11 = 88$
9	6	$9 \times 6 = 54$
10	5	$10 \times 5 = 50$
11	5	$11 \times 5 = 55$
12	1	$12 \times 1 = 12$
13	1	$13 \times 1 = 13$
Total	735	2166

Here $N = \sum f = 1343$; $\sum Xf = 2166$

Average number of children born per family is given by

$$\text{mean } (\bar{X}) = \frac{\sum Xf}{\sum x} = \frac{2166}{735} = 2.9 \text{ children}$$

Example 1.4: The following table gives the distribution of units under different heights in a certain region. Compute the mean height of the region.

Table 1.4: Height of units

Height (in metre)	200	600	1000	1400	1800	2200
Number of units	142	265	560	271	89	16

Solution: (a) By direct method

Height (in metre) (X)	Number of units (f)	$X.f$ (xf)
200	142	$200 \times 142 = 28400$
600	256	$600 \times 256 = 159000$
1000	560	$1000 \times 560 = 560000$
1400	271	$1400 \times 271 = 379400$
1800	89	$1880 \times 89 = 160200$
2200	16	$2200 \times 16 = 35200$
Total	1343	1322200

Here $N = \sum f = 1343$; $\sum Xf = 1322200$

$$\text{mean height } (\bar{X}) = \frac{\sum Xf}{\sum x} = \frac{1322200}{1343} = 984.51 \text{ metres}$$

(b) By short cut method: Let the working mean (A)= 1400

Height (in metre) (X)	X-A (d)	Number of units (f)	$d.f$ (df)
200	-1200	142	$-1200 \times 142 = 170400$
600	-800	265	$-800 \times 265 = 212000$
1400 A	0	271	$0 \times 271 = 0$
1800	+400	89	$+400 \times 89 = 35600$
2200	+800	16	$+800 \times 16 = 12800$
Total		1342	-558000

Here $N = \sum f = 1343$; $\sum fd = -558000$

$$\text{mean height } (\bar{X}) = A + \frac{\sum Xf}{N} = 1400 + \frac{(-558000)}{1343} = 984.51 \text{ metres}$$

4.3.3 Grouped Data (Continuous Frequency Distribution)

When the measurement are given in the grouped form, the mean is computed by multiplying the various mid values m_i with their respective frequencies f_i where $i = 1, 2, \dots, n$ and dividing the product total by the sum of frequencies or total number of observations. If $m_1, m_2, \dots, m_n$ are the mid values or class marks corresponding to frequencies $f_1, f_2, \dots, f_n$ then the mean is

$$\bar{X} = \frac{m_1 + m_2 f_2 + \dots + m_n f_n}{f_1 + f_2 + \dots + f_n} = \frac{\sum_{i=1}^n m_i f_i}{\sum_{i=1}^n f_i} = \frac{\sum_{i=1}^n m_i f_i}{N}$$

Where n stand for the number of groups and N denotes the total number of observations.

Short cut method

If the class groups formed by individual reading are large, the following procedure is adopted for computing arithmetic mean in a continuous series as follows:

- (i) Find out mid values of the class intervals, and assume one of the mid values (preferably in the middle of the distribution) as working mean (A).
- (ii) Calculate deviations of the mid values from the working mean ($m-A$) and denotes by d .
- (iii) Multiply each d by its respective frequency to find out fd .
- (iv) Calculate arithmetic mean $\bar{X}$ by the following formula:

$$mean = (\bar{X}) = A + \frac{\sum_{i=1}^n f_i d_i}{N}$$

Step- Deviation Method

Further if the deviations are large and intervals among consecutive mid values are equal deviations can be made small by dividing them by the class width, i.e., if $d'_i = \frac{(m_i - A)}{h}$, the formula for calculating arithmetic mean by this method is

$$mean = (\bar{X}) = A + \frac{\sum_{i=1}^n f_i d_i}{N} \times h$$

Where h is the size of the class intervals.

Example 1.5: The rainfall of 66 districts in a particular year is given below. Compute the average annual rainfall.

Rainfall (in inches)	0-10	10-20	20-30	30-40	40-50	50-60
No. of Districts	22	10	8	15	5	6

Solution: (a) By direct method

Table 1.6
Rainfall in Districts in Particular year

<i>Rainfall (in inches) (X)</i>	<i>Mid Values (m)</i>	<i>No. of Districts (f)</i>	<i>m.f (mf)</i>
0-10	$\frac{0 + 10}{2} = 5$	22	$5 \times 22 = 110$
10-20	$\frac{10 + 20}{2} = 15$	10	$15 \times 10 = 150$
20-30	$\frac{20 + 30}{2} = 25$	8	$25 \times 8 = 200$
30-40	$\frac{30 + 40}{2} = 35$	15	$35 \times 15 = 525$
40-50	$\frac{40 + 50}{2} = 45$	5	$45 \times 5 = 225$
50-60	$\frac{50 + 60}{2} = 55$	6	$55 \times 6 = 330$
Total		66	1540

$$\text{Here } N = \sum mf = 1540$$

$$\text{mean } (\bar{X}) = \frac{\sum mf}{\sum f} = \frac{1540}{66} = 23.33 \text{ inches}$$

(b) Short cut method

<i>Rainfall (in inches) (X)</i>	<i>Mid Values (m)</i>	<i>No. of Districts (f)</i>	<i>No. of Districts (f)</i>	<i>d.f (df)</i>
-------------------------------------	---------------------------	---------------------------------	---------------------------------	---------------------

0-10	5	-30	22	$-30 \times 22 = -660$
10-20	15	-20	10	$-20 \times 10 = -220$
20-30	25	-10	8	$-10 \times 8 = -80$
30-40	35(A)	0	15	$0 \times 15 = 0$
40-50	45	10	5	$10 \times 5 = 50$
50-60	55	20	6	$20 \times 6 = 120$
Total	-		66	-770

$$\text{Here } N = \sum f = 66 \quad \sum df = -770$$

$$\text{mean } (\bar{X}) = A + \frac{\sum fd}{N} = 35 + \frac{-770}{66} = 23.33 \text{ inches}$$

(a) By step- deviation Method

From the given data $h=10$

<i>Rainfall (in inches) (X)</i>	<i>Mid Values (m)</i>	<i>No. of Districts (f)</i>	<i>No. of Districts (f)</i>	<i>d.f (df)</i>
0-10	5	$\frac{5 - 35}{10} = -3$	22	$-3 \times 22 = -66$
10-20	15	$\frac{15 - 35}{10} = -2$	10	$-2 \times 10 = -20$
20-30	25	$\frac{25 - 35}{10} = -1$	8	$-1 \times 8 = -8$
30-40	35(A)	$\frac{35 - 35}{10} = 0$	15	$0 \times 15 = 0$
40-50	45	$\frac{45 - 35}{10} = +1$	5	$1 \times 5 = 5$
50-60	55	$\frac{55 - 35}{10} = +2$	6	$2 \times 6 = 12$
Total	-	66	66	-77

$$\text{Here } N = \sum f = 66 \quad \sum fd' = -77$$

$$\text{mean } (\bar{X}) = A + \frac{\sum fd'}{N} \times h = 35 + \frac{-77}{66} \times 10 = 23.33 \text{ inches}$$

Example 1.6: The amounts of money, in thousand dollars, that a sample of people contributed to political campaigns in an election are: 1,2,5,25,10,0,2,0,5,10. Answer the following questions:

- (i) What is the total amount of money contributed $\sum X_i$?
- (ii) What is the money contributed by seventh person, X_7 ?
- (iii) What is the sample size?
- (iv) Find the mean and interpret it.

Solution: First make a frequency distribution as shown below.

X_i	f_i	$f_i X_i$
0	2	0
1	1	1
2	2	4
5	2	10
10	2	20
25	1	25
Total	10	60

- (i) Total contribution, $\sum f_i X_i = 0+1+4+10+20+25 = \60 thousand
- (ii) Contribution of seventh person, $X_7 = \$2$ thousands
- (iii) Sample-size- number of case= $n=10$
- (iv) Mean $\bar{X} = \sum f_i X_i / n = 60/10 = 6$ thousand.

This value of mean shows that the average contribution to political campaigns in the \$6,000. This also tells that most of the contributions are located around a value of \$6,000. Thus the central of the distribution of campaign contribution has been located. It is not yet how well this quantity measures average.

Example 1.7: Obtain an estimate of the value of missing frequency? If the mean of the distribution is 22.5.

<i>C.I.</i>	<i>f.</i>	<i>x</i> <i>Mid value</i>	<i>fx</i>
0-10	6	5	30
10-20	9	15	135
20-30	A	25	25a
30-40	10	30	300
40-50	3	45	135
	40		600+25a

Let a be the unknown frequency

$$\bar{X} = \frac{1}{N} \sum f(x)$$

$$22.5 = \frac{1}{40} \times (600 + 25a)$$

$$22.5 = \frac{25(24 + a)}{40}$$

$$22.5 \times 40 = 25(24 + a)$$

$$\frac{900}{25} = 24 + a$$

$$36 = 24 + a$$

$$a = 36 - 24,$$

$$a = 12.$$

The missing value is 12.

Example 1.8 (On wrongly taken value)

<i>C.I.</i>	<i>f.</i>	<i>x</i> <i>Mid value</i>	<i>fx</i>
0-10	5	5	25

10-20	12	15	180
20-30	18 (15)	25	450 (375)
30-40	11	30	385
40-50	3	45	135
	49		1175

$$\bar{X} = \frac{1175}{49} = 23.98$$

Since the value 18 is wrong and the 15 is correct value than the calculation are as followed:

Present : $N = \sum f_i = 49$ then correct N is

$$(\text{Corrected}) N = 49 - 18 + 15 = 46.$$

Similarly, $(\text{Corrected}) \sum f_i = 1175 - 450 + 375 = 1100$

Then the corrected mean is

$$(\text{Corrected}) \bar{X} = 1100/46 = 23.91$$

4.3.4 Properties of Arithmetic Mean or Fundamental Theorems on Arithmetic Mean

1. First property of mean:

The sum of the deviations about the arithmetic mean equals zero. Mathematically.

$$\sum [f_i(X_i - \bar{X})] = 0$$

Proof: $\sum f_i(X_i - \bar{X}) = \sum f_i X_i - \sum f_i \bar{X}$

$$= \sum_i f_i X_i - \bar{X} \sum_i f_i = (\text{Since } \bar{X} \text{ be independent of } i)$$

$$= N\bar{X} - \bar{X}, N = 0$$

Since, $\left(\bar{X} = \frac{1}{N} \sum_i f_i, X_i \text{ where } N = \sum_i f_i, \text{ then } \sum_i f_i, X_i = N \bar{X}\right)$

This property says that if the mean is subtracted from each score, the sum of the differences will equal zero. The property results from the fact that the mean is the balance point of the distribution. The mean can be thought of as the fulcrum of a seesaw. When the score is distribution along the seesaw according to their values, the mean of the distribution occupies the position where the scores are in balance. This is known as first property of mean.

2. Second property of mean:

The sum to the squared deviations of all the scores about their arithmetic mean is minimum. That is,

$$\sum [f_i(X_i - \bar{X})]^2 = \text{minimum}$$

This is an important characteristic and is used in many areas of statistics, particularly in regression analysis.

Proof: Let us suppose that the sum of square of deviations from point (a).

$$\sum f_i(X_i - a)^2 = K$$

According to the principle of maximum, K will be minimum if,

$$\frac{\partial k}{\partial a} = 0 \text{ and } \frac{\partial^2 k}{\partial a^2} > 0$$

$$\text{Now } \frac{\partial k}{\partial a} = (-2) \sum f_i(X - a) = 0$$

$$\rightarrow \sum f_i(X - a) = 0 \rightarrow \sum f_i X_i - Na = 0$$

$$a = \frac{1}{N} \sum f_i X_i = \bar{X}$$

Again

$$\frac{\partial^2 k}{\partial a^2} = (-2) \sum f_i(-1) = 2 \sum f_i = 2N > 0$$

Hence K is minimum at $a=\bar{X}$ and

$$\sum [f_i(X_i - \bar{X})]^2 = \text{minimum}$$

Remarks, we shall see later on that,

$$\sigma^2 = \frac{1}{N} \sum f_i(X_i - \bar{X})^2$$

is a measure of dispersion.

3. Combined property of mean:

If $\bar{X}_1$ and $\bar{X}_2$ be the means of two series of sizes n_1 and n_2 respectively, then the mean of the combined series can be computed as:

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

Proof: If $\bar{X}_1$ be the mean of series $X_{11}, X_{12}, \dots, X_{1n_1}$ and $\bar{X}_2$ be the mean series $X_{21}, X_{22}, \dots, X_{2n_2}$

Then by definition,

$$\bar{X}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i = \frac{1}{n_1} (X_{11}, X_{12} \dots X_{1n_1})$$

$$\bar{X}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} X_i = \frac{1}{n_2} (X_{21}, X_{22} \dots X_{2n_2})$$

The combined series is $X_{11}, X_{12} \dots X_{1n_1}, X_{21}, X_{22} \dots X_{2n_2}$

The mean is

$$\bar{X} = \frac{1}{n_1 + n_2} [(X_{11}, X_{12} \dots X_{1n_1}) + (X_{21}, X_{22} \dots X_{2n_2})]$$

$$\bar{X} = \frac{1}{n_1 + n_2} \left[\sum_{i=1}^{n_1} X_i + \sum_{j=1}^{n_2} X_{ij} \right]$$

$$= \frac{1}{n_1 + n_2} [n_1 \bar{X}_1 + n_2 \bar{X}_2]$$

Similarly, if $\bar{X}_1, \bar{X}_2, \bar{X}_3 \dots \dots \bar{X}_k$ be the means of k series of sizes $n_1, n_2, n_3 \dots \dots n_k$ respectively then the mean $\bar{X}$ of combined series is

$$\bar{X} = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2 + \dots \dots n_n \bar{X}_n}{n_1 + n_2 + \dots \dots n_n}$$

Example 1.9: The average ages of 250 males and 210 females in a village are 41.6 and 38.5 years respectively. Find the average age combining both males and females together.

Solution: Here (Combined average) is $N_1 = 250$, $\bar{X}_1 = 41.6$ years and $N_2 = 210$, $\bar{X}_2 = 38.5$ years. Therefore

$$\begin{aligned} \bar{X} &= \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2}{n_1 + n_2} = \frac{250 \times 41.6 + 210 \times 38.5}{250 + 210} = \frac{10400 + 8085}{460} \\ &= 40.18 \text{ years.} \end{aligned}$$

4.3.5 Advantages of Mean

- (i) The mean is sensitive to the exact values of all the scores in the distribution. Since you have to add the scores to calculate the mean, a change in any of the scores will cause a change in the mean.
- (ii) Mean is very sensitive to extreme scores. If we add an extreme score (one that is very far from the mean), it would greatly disrupt the balance. The mean would have to shift a considerable distance to reestablish balance. The mean is more to extreme than is the median or the mode. This known as 2nd property of mean.
- (iii) Of the measures used for central tendency, the mean is least subject to sampling variation under most circumstances. If repeated samples are drawn from a population, the mean would vary from sample to sample. The same is true for the median and the mode. However, the mean varies less than these other measures of central tendency. This is very important in

inferential statistics and is a major reason why the mean is use in inferential statistics whenever possible.

- (iv) It takes into account all the scores in a distribution so; mean offers a good representation of the central tendency by making use of the most information.
- (v) Mean is used in many statistical formulas, making it a more widely used measure.

4.3.5 Limitations of Mean

Mean can be misleading if there are extreme values in the distribution, for example, if the distribution is skewed (asymmetrical) of the level of measurement is less than interval. Sometimes people are interested in misleading others by making use of ‘illegitimate’ statistics. The following example illustrates this point.

Example 1.10: The amounts of money that a sample of people contributed to political campaigns in the last election were, in thousands rupees; 1,2,5,25,10,0,2,0,5,10, 500. Calculate the mean.

Solution: Make a table that contains columns: X_i, f_i and $f_i X_i$ Sum all the entries in columns X_i, f_i and $f_i X_i$ to get:

$$f_i X_i = 560, \quad f_i = n = 11$$

$$\text{Mean } \bar{X} = \frac{f_i X_i}{n} = \frac{(1 + 2 + 5 + 25 + \dots + \dots + 10 + 500)}{11}$$

$$= \frac{560}{11} = 50.91 \quad \text{or} \quad \text{Rs. } 50,910$$

A mean of Rs. 50.91 thousand suggests that the typical contribution was Rs. 50,910. We notice that the mean in this example is not at all a measure of central tendency. All but one person contributed less than the mean. This is due to the presence of an extreme contributor who contributed Rs. 500,000. This high value inflates the mean making it a misleading statistic in each situation.

4.4 Geometric Mean

In case of finding the average or rates and ratios, geometric mean is more useful measure than others, e.g. in finding the population increase simple and compound interests etc.

Case 1: In case of ungrouped data it is obtained by multiplying together all the values of the variable and extracting the relevant root of the product. i.e. if $(X_1, X_2, X_3 \dots X_n)^{\frac{1}{n}}$ and n values of a variable under then the geometric mean (G.M.) is computed as:

$$G. M. = (X_1, X_2, X_3 \dots X_n)^{\frac{1}{n}} ; (X_1 > 0)$$

To facilitate the computation one can make the use of logarithms as:

$$\begin{aligned} \log(G. M.) &= \frac{1}{n} \log(X_1, X_2, X_3 \dots X_n). \\ &= \frac{1}{n} [\log X_1 + \log X_2 + \log X_3 \dots + \log X_n] \\ &= \frac{1}{n} \left[\sum_{i=1}^n \log X_i \right] \end{aligned}$$

$$\text{So } G. M. = \text{Antilog } \frac{1}{n} \left[\sum_{i=1}^n \log X_i \right]$$

Thus the logarithm of the geometric mean is equal to the arithmetic mean of the logarithms of individuals' measurements.

Case 2: In case of grouped data if $f_1, f_2, f_3 \dots f_n$ be the frequencies corresponding to the individual values $X_1, X_2, X_3 \dots X_n$ then G.M. is computed as:

$$G. M. = (X_1^{f_1} \cdot X_2^{f_2} \cdot X_3^{f_3} \dots X_n^{f_n})^{\frac{1}{\sum_{i=1}^n f_i}} ; (X_1 > 0)$$

$$\text{or } \log G. M. = \frac{1}{\sum_{i=1}^n f_i} [f_1 \log X_1 + f_2 \log X_2 + f_3 \log X_3 \dots + f_n \log X_n]$$

$$= \frac{1}{\sum_{i=1}^n f_i} \left[\sum_{i=1}^n f_i \log X_i \right]$$

$$\text{So } G.M. = \text{Antilog} \left\{ \frac{1}{\sum_{i=1}^n f_i} \left[\sum_{i=1}^n f_i \log X_i \right] \right\}$$

Here $\log G$ is the weighted mean of $\log X_i$'s with weights $f_1, f_2, \dots, f_n$.

Additive Property of Geometric Mean:

If G_1 and G_2 are the geometric means of two series with respective sizes n_1 and n_2 , the combined Geometric mean G is

$$\log G = \frac{n_1 \log G_1 + n_2 \log G_2}{n_1 + n_2}$$

Proof: Suppose that $X_{11}, X_{12}, \dots, X_{1n}$ be the first and $X_{21}, X_{22}, \dots, X_{2n_2}$ be the 2nd series with sizes n_1 and n_2 . Then

$$G_1 = \left[\prod_{i=1}^{n_1} x_{1i} \right]^{1/n_1} \quad \text{and} \quad G_2 = \left[\prod_{j=1}^{n_2} x_{2j} \right]^{1/n_2}$$

$$\log G_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \log_e x_{1i}, \quad \log G_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} \log x_{2j}^{n_2}$$

G.M. of combined series $X_{11}, X_{12}, \dots, X_{1n}, X_{21}, X_{22}, \dots, X_{2n_2}$ is,

$$G = \left[\prod_{k=1}^{n_1+n_2} x_{k^2}^{\frac{1}{n_1+n_2}} \right] \rightarrow \log G = \frac{1}{n_1 + n_2} \left[\sum_{i=1}^{n_1} \log x_{1i}^{n_1} + \sum_{j=1}^{n_2} \log x_{2j}^{n_2} \right]$$

$$\rightarrow \log_e G = \frac{1}{n_1 + n_2} [n_1 \log_e G_1 + n_2 \log_e G_2]$$

Example 1.11: The monthly average temperature of a situation for five months is given as 16.2, 23.4, 20.6, 33.4 and 16.4 degree centigrade. Find the mean temperature of the station.

Solution: $G.M. = (16.2 \times 23.4 \times 20.6 \times 33.4 \times 16.4)^{1/2}$

or $\log G.M. = \frac{1}{5} [\log 16.2 + \dots + \log 16.4]$

$$= \frac{1}{5} [1.2095 + 1.3692 + 1.3139 + 1.5238 + 1.2148]$$

$$= \frac{1}{5} \times 6.6312 = 1.3262$$

So G.M. = Antilog [1.3262] = 21.1934 degree centigrade.

Example 1.12: From the following data, calculate the G.M.

Class group	0-10	10-20	20-30	30-40	40-50
No. of observations	14	23	27	21	15

Calculation:

<i>Rainfall (in inches)</i>	<i>Mid Values (x)</i>	<i>log (x)</i>	<i>No. of observations (f)</i>	<i>f log x</i>
0-10	5	0.69897	14	9.78558
10-20	15	1.176.9	23	27.05007
20-30	25	1.39794	27	37.74438
30-40	35	1.54407	21	32.42547
40-50	45	1.65321	15	24.79815
Total	-		100	131.80365

Here $\sum f = N = 100$, $\sum f \log x = 131.80365$.

$$G.M. = \text{Antilog} = \left\{ \frac{1}{\sum f} [\sum f \log x] \right\}$$

$$\text{Antilog} = \left[\frac{131.80365}{100} \right] = \text{Antilog} [1.318036] = 20.7987$$

Use of Geometric Mean in Computing Rate of Growth

Geometric mean provides a satisfactory measure for computing rate of growth of population phenomena, special the phenomena which grow at geometric progression.

The formula is $P_1 = P_0(1 + r)^t$ or $r = \left(\frac{P_1}{P_0} - 1\right)^{\frac{1}{t}}$

Where, P_1 = the value of variable at the end of the period, i.e. at time t.

P_0 = the value of the variable at the beginning

r = the rate of growth per unit of time

t = number of units of time.

Example 1.13: Population of India in 1961 and 1971 were 43.9 and 54.8 crores. Find the rate of increase.

Solution: Here $P_1 = 54.8$; $P_0 = 43.9$, $t = 10$ years; $r = ?$

$$r = \left(\frac{P_1}{P_0} - 1\right)^{\frac{1}{t}} = \left(\frac{54.8}{43.9} - 1\right)^{\frac{1}{10}}$$

$$\text{or } \log r = \frac{1}{10} \log \left(\frac{54.8}{43.9} - 1\right) = \frac{1}{10} (.248292)$$

$$r = \text{Antilog} (.0248292) = 1.05884.$$

4.5 Harmonic Mean

In problems such as work time and rate where the amount of work is held constant an average rate is required, the harmonic mean (HM) is utilized. It is defined as the reciprocal of the arithmetic mean of the reciprocals of the given individual readings i.e. H.M. of $X_1, X_2, \dots, X_n$ is defined as:

$$H.M. = \frac{n}{\frac{1}{X_1} + \frac{1}{X_2} + \dots + \frac{1}{X_n}} \quad (X_i > 0)$$

$$= \frac{n}{\sum_{i=1}^n \frac{1}{X_i}}, \text{ Where } n \text{ is the number of observations.}$$

Check your progress

1. Explain the geometric mean.
2. What do you mean by harmonic mean?

Example 1.14: The H.M. of 2,4,6 is $H.M. = \frac{3}{\frac{1}{2} + \frac{1}{4} + \frac{1}{6}} = \frac{36}{11} = 3.37$.

In case of frequency distribution (grouped data), if $f_1, f_2, \dots, f_n$ be the frequencies corresponding to $X_1, X_2, \dots, X_n$ then H.M. is computed as:

$$H.M. = \frac{f_1 + f_2 + f_3 + \dots + f_n}{\frac{f_1}{X_1} + \frac{f_2}{X_2} + \dots + \frac{f_n}{X_n}} \quad (X_i > 0)$$

$$= \frac{\sum_{i=1}^n f_i}{\sum_{i=1}^n \frac{f_i}{X_i}}$$

Example 1.15: Compute H.M. for the data given in example 9.

Calculation:

<i>Class group (X)</i>	<i>Mid Values (m)</i>	<i>No. of observations (f)</i>	<i>f / m</i>
0-10	5	14	2.80
10-20	15	23	1.53
20-30	25	27	1.08
30-40	35	21	0.60
40-50	45	15	0.33
Total	-	100	6.34

Here $\sum f = N = 100$, $\sum f / m = 6.34$

$$H.M. = \frac{\Sigma f}{\Sigma\left(\frac{f}{m}\right)} = \frac{100}{6.34} = 15.77.$$

4.6 Exercises

Q.1. Data are collected on the weekly expenditures of a sample of urban households on food. The data obtained from diaries kept by each household, are grouped by number of members of the household. The expenditures were as follows:

1 member: 67 62 168 128 131 118 80 53 99 68 76 55 84 77 70 140 84 65 67 183

2 member: 129 116 122 70 141 102 120 75 114 81 106 95 94 98 85 81 67 69 119 105 94 94 92

3 member: 79 82 99 142 171 82 145 94 86 85 100 191 116 100 125 116.

4 member: 139 111 251 106 93 99 155 132 158 62 114 129 108 91.

5 or more members: 121 128 129 140 206 111 104 109 135 136.

For each number of members calculate the mean, median, mode, 25th percentile, 50th percentile and 75 the percentile. Interpret each statistic.

Q.2. An instructor gives a quiz with three questions, each worth 1 point; 40% of the class scored 3 points, 30% scored 2 points, 20% scored 1. And 10% scored 0;

Score	Percentage
3	40
2	30
1	20
0	10

- (i) if there were ten people in the class, what would the average score be?
- (ii) If there were twenty people in the class, what would the average score be/
- (iii) Suppose you are not told the number of people in the class. Can you still figure out the average score? Explain.

Q.3. In 1989, Governor Brown of California proposed that all state employees be given a flat raise of \$70 a month. What would this do to the average monthly salary of state employees? What would

a 5% increase in the salaries, across the board do to the average monthly salary? What will the doubling of salaries do to the mean salary?

4.7 Summary

Various measures of central tendency have been defined in this unit. There is found a tendency in the data to cluster around a central value. This value is known as measure of central tendency. These are mean, median and mode. Mean is obtained by dividing the sum of observations by number of observations. Median is that variate value which divides the given data or frequency distribution in two equal halves. Mode is that variate value which occurs most frequently i.e. for which the frequency is maximum. Mean, median and mode approximately satisfy the relation.

$$\text{Mean-mode} = 3 (\text{mean}-\text{median})$$

4.8 Further Readings

1. Goon A.N., Gupta M.K. & Das Gupta B (1987) *Fundamentals of Statistics Vol. I* The World Press Pvt. Ltd., Kolkata.
2. Yule G.U. and Kendall, M.G.: *An Introduction to the Theory of Statistics*, Charles Griffin and Company Ltd.

Unit-5: Measures of Central Tendency-II

Structure

5.1 Introduction

5.2 Objectives

5.3 Median

5.3.1 Calculation of Median (Ungrouped Data)

5.3.2 Calculation of Median (Grouped Data)

5.3.3 Calculation of Median (Graphic Method)

5.3.4 Advantages of Median

5.3.5 Disadvantages of Median

5.4 Mode

5.4.1 Calculation of Mode (Ungrouped Data)

5.4.2 Discrete Series (Grouped Data)

5.4.3 Continuous Series (Grouped)

5.5 Percentile, Deciles and Quartiles to Measurement of Location

5.6 Percentile Score from Given Percentile Rank

5.7 Choice of Measurement

5.8 Exercises

5.9 Summary

5.1 Introduction

Statistical methodology encompasses a wide range of methods involved in collecting, processing, condensing, and analyzing data. One key aspect is the tendency of observations to cluster around a central value, known as "central tendency." Measures of central tendency provide a single point around which observations tend to cluster. This central value, often called an "average," is useful for determining the distribution's location.

This unit covers various measures of central tendency, including situations where they are calculated for both ungrouped and grouped data.

5.2 Objectives

After studying this unit, you will be able to-

- Understand the meaning of median
- Compute common measures of central tendency, i.e. median and mode.
- Compute the various measures of partitions of data such as quartiles deciles and percentiles.
- Understand how to choose proper measure of central tendency.

5.3 Median

Median is another important and useful measure of central tendency. It has connotation of the middle most or most central value of a set of measurements. It is usually defined as the value which divides a distribution in such a manner that the number of items below it is equal to the number of items above it. The median is thus a positional average. It is better indication of central tendency when one or two of the peripheral readings are too large or too small because they give the wrong idea of the average when mean is computed. Median is that variate value of the data or frequency distribution which divides it in two equal halves.

5.3.1 Calculation of Median (Ungrouped data)

Case 1 (n is odd): In case of ungrouped data when the number of observations are odd, then median is the middle value after the measurements have been arranged in ascending or descending order of magnitude, i.e. if there are n number of measurements and measurements are arranged in ascending or descending order of magnitude, the median of the measurements is $\left(\frac{n+1}{2}\right)^{th}$ measurement where n is an odd number.

Case 2 (n is even): If the number of observations are even, median is defined as the mean of the two middle observations are arranged in ascending or descending order to magnitude i.e.

$$median = \frac{\left(\frac{n}{2}\right)^{th} value + \left(\frac{n+1}{2}\right)^{th} value}{2}$$

Example 1.16: Calculate median for the following data:

(a) 68, 62, 75, 82, 68, 71, 68, 71, 62, 68, 74, 59, 74, 68, 60, 71, 59, 73, 73, 58.

(b) 200, 150, 260, 285, 380, 305, 4989, 307, 1280, 233, 403

Solution (a) To compute the median first we arrange the values in ascending order of magnitude as: 58, 59, 59, 60, 62, 62, **68, 68, 68, 68**, 71, 71, 71, 73, 73, 74, 74, 75, 82.

The number of observation n is even in this case, i.e., n=20.

$$\text{So } median = \frac{\left(\frac{n}{2}\right)^{th} \text{ value} + \left(\frac{n+1}{2}\right)^{th} \text{ value}}{2} = \frac{10th \text{ value} + 11th \text{ value}}{2} = \frac{68+68}{2} = 68$$

(b) Let us first arrange the values in ascending order of magnitude as:

150, 200, 233, 260, 285, 305, 307, 380, 403, 1280, 4989

The number of observations n in this case is odd, i.e., $n=11$ so the median is the $\left(\frac{n+1}{2}\right)^{th}$ value i.e., $\left(\frac{11+1}{2}\right)^{th}$ or 6th value of the observation and thus underlined value. i.e., 305 is the median.

5.3.2 Calculation of Median (Grouped Data)

In case of discrete frequency distribution median can be obtained with the help of cumulative frequencies as follows:

- (i) First find $N/2$ where $N = \sum f$
- (ii) Find the cumulative frequency just greater than $N/2$.
- (iii) Corresponding value of X (i.e., of variable) is median.

Example 1.17: Calculating the median height from the data given in example 1.4.

Calculation

Height (in metre) (X)	Number of units (f)	Cumulative frequency (F)
200	142	142
600	256	407
1000	560	967
1400	271	1238
1800	89	1327
2200	16	1343
Total	1343	

Here $f=N= 1343$; $\frac{N}{2} = 671.5$

The cumulative frequency just greater than 671.5 is 967 and corresponding to this cumulative frequency, the value of X is 1000 and thus the median height is 1000 meters.

Median (Continuous Grouped Data): Median for such distribution is computed by the following formula

$$Median (Md) = l_m + \frac{\left(\frac{n}{2} - f_e\right)}{f_m} \cdot h$$

Where l_m is the lower limit f_m is the frequency of the median class, f_e is the cumulative frequency of the class, preceding the median class and h is the width of the median class and $N = \sum f_i$

Example 1.18: Calculate mediana for the following grouped data.

Interval	35-45	45-55	55-65	65-75	75-85
Frequency	2	3	5	1	1
Cum. Freq.	2	5	10	11	12

The cum. Freq. is computed from the given freq. dist.

The mediana position = $(n+1)/2 = (12+1)/2 = 6.5$

Median lies between observation 55 and 65. Both of these observations fall in category 3, i.e., in class (35-65) with cumulative frequency of 10. Therefore,

$$Median (Md) = l_m + \frac{\left(\frac{n}{2} - f_e\right)}{f_m} \cdot h$$

Where $l_m = 55$, $f_m = 5$, $h=10$, $f_e = 5$ $N=12$

$$Median (Md) = 55 + \frac{[(12/2) - 5]}{5} \times 10 = 55 + 2 = 57$$

Example 1.19: The following table gives the size of land holding of families in a village. Find out the median holding size.

Area of land (in acres)	5-9	10-14	15-19	20-24	25-29	30-34
No. of families	20	35	150	70	44	38

Solution: Since the class groups are given in the discrete form hence we first have to convert it into continuous form by adding .5 to the upper limits and subtracting .5 from the lower limits as given in column 2 below:

Area of land (in acres) (X)	Area of land (in acres) Continuous from	No. of families (f)	Cumulative frequency (f)
5-9	4.5-9.5	20	20
10-14	9.5-14.5	35	55
15-19	14.5-19.5	150	205
20-24	19.5-24.5	70	275
25-29	24.5-29.5	44	319
30-34	29.5-34.5	38	357
Total		357	

Here $N/2 = 178.5$, the cumulative frequency just greater than 178.5 is 275 and the corresponding class group is the median class. For this median class, we have

Where $l_m = 14.5$, $f_c = 55$, $h = 5$, $f_e = 150$

$$\text{Median (Md)} = 14.5 + \frac{[178.5 - 55]}{150} \times 5 = 14.5 + 4.12 = 18.62 \text{ acres.}$$

5.3.3 Calculation of Median by (Graphical Method)

One of the methods to compute median is the graphical method. In this case take the class intervals (or the individual readings) on the axis of X and plot the corresponding cumulative frequencies on the axis of Y against the upper limit of the class interval (or against the variate value in case of discrete frequency distribution). The Curve obtained by joining the points by means of free hand drawing is the cumulative frequency curve or ogive. For the calculation of median take a point on

the axis of Y that is equivalent to $N/2$ and from this point draw a line parallel to X-axis. This line will cut the curve and from the cutting point draw a line perpendicular on X axis. This distance from origin to the point at which the perpendicular line cuts the X axis is the value of median.

Example 1.20: Find out the median rainfall from the distribution given in example 1.13 by the graphical methods:

Figure 3.1 shows the cumulative frequency curve formed between the upper limits of classes and corresponding cumulative frequencies. The point $N/2$ is shown on Y axis and the dotted line parallel to X-axis cuts the cumulative curve at C. Perpendicular line cuts the X-axis at M. The distance OM is the median value.

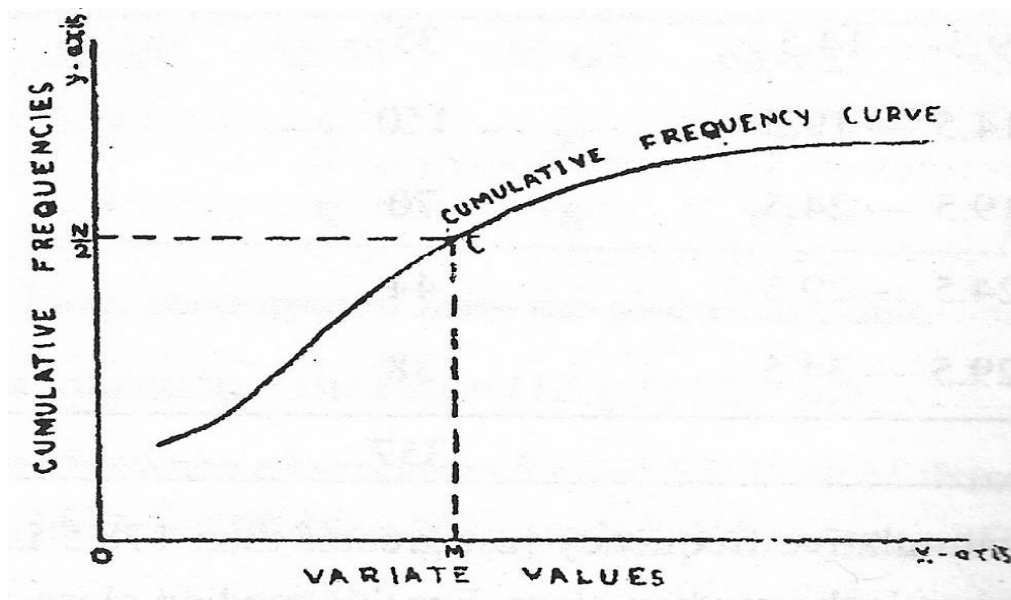


Fig. 3.1 Cumulative Frequency Curve

Check your progress

1. Explain the median.
2. How to calculate the median by graphical method.

5.3.4 Advantages of Median

An important advantage of median is that it is less sensitive than the mean to extreme scores. For skewed data, the median is a better choice because it is usually not affected by a few outlier. Median is also a desirable measure when the distribution has to be truncated for some reasons. If the purpose is to describe the central tendency of a set of scores, the median is preferable to other measures. It gives an undistorted picture of central tendency whether the data are skewed or not.

5.3.5 Disadvantages of Median

Under usual circumstances the median is more vulnerable to sampling variable than the arithmetic mean. This makes median less stable than the mean from sample to sample and therefore it is not very useful in inferential statistics. For ordinal data median also ignores the actual values of observations and simply takes into account their positions.

5.7 Mode

Mode or modal value of a distribution is that the value which occurs most frequency, for example, at any station the average number of occurrences for thunder storms or days with snowfall wind direction, etc are the most realistically presented by modal value. In case of frequency distribution, the mode is that value which has maximum frequency. If two or more observations occur the same number of times, then there is more than one mode and the distribution is called multi-model.

Types of Measures of Central Tendency

The measures of central tendency or average are of different types, but the most common in use are of three types:

1. Mean
2. Median
3. Mode

The mean is further classified as:

- (i) Arithmetic mean
- (ii) Geometric mean
- (iii) Harmonic mean

Since each one of the above measures of central tendency has its own individual characteristics and properties, a decision must always be made as to which would be the most appropriate and useful in view of the nature of the statistical data and purpose of the inquiry. The qualities desired in a measure should be (a) rigidly defined, (b) easily computed, (c) capable of a simple interpretation (d) not unduly influenced by one or two extremely large or small values and (e) likely to fluctuate relatively little from one random sample to another (of the same size and from the same population). However, the decision about which of the three measures of central tendency to use will be clear after learning the computation of each one. A few general considerations in choosing a measure of central tendency are: (i) the purpose of research- what characteristic of the data are of interest; (ii) the level of measurement of the data- nominal ordinal, interval or ratio level; (iii) the shape of the frequency distribution as indicated by a graph- symmetric or skewed; (iv) level of expertise of the researcher and the audience- what can you accomplish and what your audience is able to understand.

5.7.1 Calculation of Mode (Ungrouped Data)

Mode is defined as that variate value of the data or the frequency distribution which occurs most frequently.

The mode in a series of individual measurement can be located either of two ways.

- (i) Data should first be an array so that repetition of a value can be identified and quickly counted, the value of that item which occurs most of the times is the modal value.
- (ii) Data should be converted into a discrete series.

Example 1.21: Find the modal temperature value from the values given in example 1.14

Solution (i) Putting data in array as:

58, 59, 59, 60, 62, 62, **68, 68, 68, 68**, 71, 71, 71, 73, 73, 74, 74, 75, 82.

Here mode = 68⁰ F.

(ii) Discrete series (converted to frequency distribution form)

Variable (X)	58	59	60	62	68	71	73	74	75	82
Frequency (f)	1	2	1	2	5	3	2	2	1	1

Here the value 68 occurs the maximum number of times, hence it is mode.

5.7.2 Discrete Series (Grouped Data)

In case of discrete frequency distribution, mode can be located by inspection of the distribution alone. The size having the maximum frequency will be reckoned as mode.

Example 1.22: Computed the modal size of children born per family in the locality from the data given in example 1.3.

Solution: The highest size of frequency in the given distribution is 154 and corresponding to this frequency the number of children born per family is 2. Hence the modal size of children born per family in the locality is 2.

5.7.3 Continuous Series (Grouped)

- (i) Determine the modal class interval. It is the data class interval with the maximum number of frequencies in it. This can be found out by just observing the series.
- (ii) Determine the value of mode by applying the following formula:

$$Mode = L + \left(\frac{f - f_p}{2f - f_p - f_s} \right) h$$

Where, L is the lower limit of the modal class; f is the frequency of the modal class; f_p is the frequency of the class preceding the modal class ; f_s is the frequency of the class succeeding the modal class and h is the class width of the modal class.

Example 1.23: Compute the modal agricultural holding of the village from the data given in example 1.19.

Solution: The maximum frequency 'f' in the distribution is 150 which corresponds to class group 15-19, i.e., 14.5-19.5 in continuous case (see column 2, example 14). Hence modal class is 14.5-19.5. Now mode is computed as: $Mode = L + \left(\frac{f - f_P}{2f - f_P - f_S} \right) \times h$

Here $L = 14.5$, $f = 150$, $f_P = 35$, $f_S = 70$ and $h = 5$.

$$\begin{aligned} Mode &= 14.5 + \left(\frac{150 - 35}{150 \times 2 - 35 - 70} \right) \times 5 \\ &= 14.5 + \frac{115}{195} \times 5 = 17.45 \text{ acres.} \end{aligned}$$

Sometimes mode is also computed with the help of mean and median. For a symmetrical distribution mean, median and mode coincide and if the distribution is moderately asymmetrical, the mean, median and mode are approximately related by the formula:

$$Mode = 3 \text{ Median} - 2 \text{ Mean}$$

Example 1.24: If the mean and median of a moderately asymmetrical series are 12.9 and 12.1 respectively, what would be its most probable mode?

Solution: Mean = 12.9, Median = 12.1, Mode ?

$$Mode = 3 \text{ Median} - 2 \text{ Mean}$$

$$= 3 \times 12.1 - 2 \times 12.9$$

$$= 36.3 - 25.8$$

$$= 10.5.$$

5.8 Percentiles, Deciles and Quartiles

In general, the term 'fractile' refers to a score that divides a distribution into fractional parts. Percentiles is the most commonly used fractile. Other fractiles are deciles, quartiles etc.

Percentiles, deciles and quartiles are used as measures of location- to locate the positions of a score relative to other to scores in a distribution. These measures are close relative to median.

Percentile is a score below which a certain percentage of score falls. For example, a student falling at the ninety- first percentage on an examination means that 91 percent of the students had score lower than his/her.

Percentile score is the raw corresponding to a percentile rank, where percentile rank is the rank (from 0 to 100) at which a particular raw score falls. A percentile rank tells us the percentage of scores falling below a score. Such a score is referred to as percentile.

Percentile divide the distribution into 100 portions of equal size. For example, sixty-fifth percentile is the score below which 65% of the cases fall. Similarly, deciles divide the distribution into 10 portions of equal size. For example, third deciles is the score below which 30% of the score fall. Calculating the third decile is equivalent to calculating the 30th percentile. Quartiles divide the distribution into 4 portions of equal size. For example, second quartile is the point or score below which 50% of the scores fall. Calculating the 2nd quartile is equivalent to the 50th percentile.

Each percentile equals 1 % of a distribution : $1 \times 100 = 100$

Each decile equals 10% of a distribution: $10 \times 10 = 100$

Each quartile equal 25% of a distribution: $25 \times 25 = 100$

Notice that the percentile ranks and percentile scores can also be read directly from cumulative frequency graphs or the cumulative frequency column of a frequency table.

5.8.1 Percentile Score from Given Percentile Rank

Percentiles can be computed using mathematical formula. Make a frequency distribution and locate the interval in which the percentile of interest belongs. This can be done by using the column of cumulative percentage frequencies. Suppose one is interested in finding the value of the pth percentile. Referring to the column of cumulative percentage frequencies, locate the class interval

that contains the p^{th} percentile. Then use the following formula to determine the approximate value of the p^{th} percentile.

$$P_p = l_p + \frac{\frac{pn}{100} - F_p}{f_p} \times w_p$$

Where F_p = cumulative frequency upto but not including the p^{th} percentile category

f_p = number of cases in the interval containing p^{th} percentile

l_p = lower limit of interval containing p^{th} percentile

w_p = width of interval containing p^{th} percentile.

p = percentile rank

n = sample size

To calculate Percentile Rank Given a Percentile Score

$$p \times \frac{F_x + [(X - l_x)/w_x]f_x}{n} \times 100$$

Where F_x = Cumulative frequency up to but into including the interval containing the X ;

X = given raw scores

w_x = width of interval containing X

l_x = lower limit of interval containing X

f_x = frequency in interval containing X

n = total number of scores or cases.

Example 1.25: Consider the following distribution of number of prisoner's arrest for 50 inmates at a state prison.

Interval	True limits	f_i	cf_i
0-2	-.5-2.5	0	0
3-5	2.5-5.5	17	17
6-8	5.5-8.5	15	32
9-11	8.5-11.5	8	40

12-14	11.5-14.5	4	44
15-17	14.5-17.5	3	47
18-20	17.5-20.5	1	48
21-23	20.5-23.5	1	49
24-26	23.5-26.5	0	49
27-29	26.5-29.5	1	50
Total		50	

- (i) Find the percentile rank of a prisoner who has been arrested 6 times.
- (ii) How many times a prison has been arrested in order to be at the (a) 2nd quartile (b) 3rd quartile?

Solution:

- (i) Percentile rank p_x , from a given score X where $X=6$.

$$p_x = \frac{F_x + [(X - l_x)/w_x]f_x}{n} \times 100$$

Where $X = 6$ (it fall in interval 5.5-8.5); $F_6 = 17$; $w_6 = 3$; $l_6 = 5.5$; $f_6 = 15$; $n = 50$

$$p_x = \frac{17 + [(6 - 5.5)/3]15}{50} \times 100 = \left[\frac{17 + 2.5}{50} \right] \times 100 = 39$$

A percentile rank of 39 means 39% of the prisoner were arrested 6 times or less.

- (ii) (a) Percentile score X_p for a Given percentile rank p of 50.

$$X_p = l_p + \frac{\frac{pn}{100} - F_p}{f_n} \times w_p$$

Where $p = 50$ (2nd quartile is equivalent to 50th percentile). Also notice that the 50th percentile falls in the interval 5.5-5.8 since 50% of the score of the score are in or below this interval.

$F_{50} = 17$; $f_{50} = 15$; $l_{50} = 5.5$; $w_{50} = 3$

$$X_{50} = 5.5 + \frac{\frac{50 \times 50}{100} - 17}{15} \times 3 = 5.5 + \left(\frac{8}{15} \right) (3) = 5.5 + 1.6 = 7.1$$

A percentile of 7.1 means a prisoner should have been arrested about 7 times to be at the 2nd quartile or the 50th percentile.

(ii) (b) Percentile score X_p for a given percentile Rank p of 75;

$$X_p = l_p + \frac{\frac{pn}{100} - F_p}{f_n} \times w_p$$

Where $p = 75$ (3rd quartile is equivalent to 75th percentile). Also notice that the 75th percentile falls in the interval 8.5-11.5 since 75% of the score of the score are in or below this interval.

$F_{75}=32$; $f_{75}=8$; $l_{75}=8.5$; $w_{75}=3$; $n=50$

$$X_{75} = 8.5 + \frac{\frac{75 \times 50}{100} - 32}{8} \times 3 = 8.5 + \left(\frac{5.5}{8}\right)(3) = 10.56$$

A percentile score of 10.56 means a prisoner should have been arrested about 11 times to be at the 3rd quartile or the 75th percentile. In other words, 75% of the prisoners were arrested less than 11 times.

5.9 Choice of Measure of Average

Following are a few important criteria in choosing a measure of average;

- (1) If there is a specific purpose or goal in mind, choose a measure of central tendency that will help to achieve that goal. The measure chosen may or may not be appropriate. An inappropriate measure (calculated in violation of its assumptions) may serve the purpose better than an appropriate measure can.
- (2) If the variable is nominal, only mode can be calculated, and thus the choice is simple. If the variable is ordinal, both mode and median can be calculated. But median is a better measure because it makes use of more information. If the variable is interval or ratio level all three measures can be calculated. In that case if the distribution is normal (symmetric and bell shaped), all three measures will have the same value. But if the distribution is skewed, median is a better measure of central tendency because it ignores the extreme scores responsible for causing the skewness.

Relationship Among Measures of Average

For symmetrically shaped distribution: Mean = Median=Mode

For positively- skewed distribution: Mean > Median> Mode

For negatively skewed distribution: Mean < Median< Mode.

5.10 Exercises

Q.1. A politician charges the opposition political party with spending an average of over Rs. 100,000 for its candidates from the state and that this is an outrage sum for a party to spend on its candidates, specially on candidates for state senator and state representative. The campaign spending figures for the party are:

Office	No. of Candidates	Average amount spent (Rs)
U.S. Senator	2	1,000,000
U.S. Congressman	16	400,000
Governor	1	800,000
State Senator	50	35,000
State rep.	50	23,000

Calculate mean, median and mode for campaign spending. Is the politician's criticism valid? Explain

Q.2. In a corporation, a very small group of employees has extremely high salaries while the majority of employees receive much lower salaries. If you were the bargaining agent for the employees, what measure of average would you calculate to illustrate the low pay level and why "if you were the employer what kind of average would you use to demonstrate a high pay level, and why?

Q.3. Which measures of central tendency are appropriate for each of the following variables? If several can be calculated, indicate which makes most use of the available information. Comment briefly on each.

- (a) Number of siblings
- (b) Political party affiliation (c) Satisfactory with family
- (d) Vacation days per year (e) Type of car driven.

5.11 Summary

Various measures of central tendency have been defined in this unit. There is found a tendency in the data to cluster around a central value. This value is known as measure of central tendency. These are mean, median and mode. Mean is obtained by dividing the sum of observations by number of observations. Median is that variate value which divides the given data or frequency distribution in two equal halves. Mode is that variate value which occurs most frequently i.e. for which the frequency is maximum. Mean, median and mode approximately satisfy the relation.

$$\text{Mean-mode} = 3 (\text{mean-median})$$

5.12 Further Readings

1. Goon A.N., Gupta M.K. & Das Gupta B (1987) *Fundamentals of Statistics Vol. I* The World Press Pvt. Ltd., Kolkata.
2. Yule G.U. and Kendall, M.G.: *An Introduction to the Theory of Statistics*, Charles Griffin and Company Ltd.

Unit-6: Measures of Dispersion

Structure

6.1 Introduction

6.2 Objectives

6.3 Types of Measures of Dispersion

6.4 Range

6.5 Mean Deviation

6.6 Variance and Standard Deviation

**6.6 Relationship between Measures of Central Tendency and Measures
of Dispersion**

6.7 Coefficient of Variation (CV)

6.8 Exercises

6.9 Summary

6.10 Further Readings

6.1 Introduction

The objective of numerical description is to obtain a set of measures that will create a mental reconstruction of the frequency distribution of the data. Now we know something about averages; let us turn to another type of measures, called the measures of dispersion. These measures summarize how spread out scores are. It is useful to know how similar or dissimilar scores are from the average score and from one another. We might like to know if scores cluster, they are more homogeneous. If scores are spread out widely then they are more heterogeneous. The measures of central tendency as discussed above only located the central of a distribution but tell nothing about the degree of variability. In order to study the dispersion or variability in a distribution we need alternative measures called the measures of dispersion.

Measuring dispersion is important for two reasons. First quantifying the dispersion in the data is required by many of the statistical inference tests that will be discussed later. Second dispersion of a distribution, in conjunction with the central tendency, more completely describes the distribution. Whereas measures of central tendency are quantification of the average value of the distribution measures of dispersion are quantification of the extent of dispersion.

This can be illustrated with the following sample:

Suppose 20 students, 10 in a sociology class and 10 in a statistics class, are asked how many hours of TV they watched last week. Their answers are as follows:

Sociology (hours)	4	4	5	5	5	5	5	5	6	6
Statistics (hours)	0	0	0	0	0	3	10	10	12	15

Mean number of hours for sociology students = $(4+4+5+\dots+6+6)/10=5$.

Mean number of hours for statistics students = $(0+0+\dots+12+15)/10=5$.

The average number of hours for two groups students is same- 5 hours per week. However, there is a difference in the way values are distributed in the two distributions. All of the sociology students watched TV between 4 and 6 hours during the week and there is little variation in the hours from student to student. These statistics students however differ from each other much more. Some seem to have devoted themselves entirely away from TV, while others watched a lot of TV. Thus several distributions may have the same mean but differ from each other in the way scores are distributed.

In central tendency is thought of as the point that best represents a central score in a distribution, the dispersion presents the other side of the coin. Dispersion reflects the “goodness” or “poorness” of central tendency as a representation of all the scores in a distribution.

6.2 Objectives

After studying this unit, you will be able to understand:

- Methods of computing various measures of dispersion;
- The advantages as well as limitations of each of these measures;
- The relationship between measures of central tendency and measures of dispersion.
- The coefficient of variation as a measure for comparing two distributions.

6.3 Types of Measures of Dispersion

Dispersion is defined as the degree to which scores deviate from the central tendency (usually the mean) of the distribution. The statistical techniques that quantify this dispersion in a distribution are called measures of dispersion. Most commonly used measures of dispersion are range, average deviations, variance and standard deviation.

6.4 Range

Range is defined as the difference between the highest and the lowest scores in a distribution. Symbolically, $R = X_{\max} - X_{\min}$

Where R is the range, $X_{\max}$ is the highest score, $X_{\min}$ is the lowest score.

A large value of range indicates greater dispersion and a small value of range indicates lesser dispersion among the scores. Minimum value that range can achieve is 0 and the maximum is infinity. If all the scores are the same, R will have a value of 0 and hence there is no dispersion.

Example 2: Find range for value: 87, 92, 47, 58, 87, 62, 73, 73, 61.

Solution: It is always a good idea to first rank the observation in ascending or descending order. In an ascending order the scores are: 47, 58, 61, 62, 73, 73, 87, 87, 92. A visual examination shows that $X_{\max} = 92$; $X_{\min} = 47$. Therefore, $R = 92 - 47 = 45$.

6.4.1 Advantages of Range

- (i) Range gives a quick identification of dispersion. It can be a good measure if there are no outliers in the data that means the distribution is not skewed.
- (ii) Range is easy to compute and interpret. For variables measured at an ordinal scale. Range is the only measure which is technically meaningful.
- (iii) If the data are to be presented to a relatively unsophisticated audience, the range may be the only measure of dispersion that will be readily understood.

6.4.2 Disadvantages of Range

- (i) Calculation of range is based only on two extreme scores, the minimum and the maximum. The rest of the data are ignored.
- (ii) Range tells nothing about the dispersion among intermediate scores.

- (iii) Range is greatly affected by outliers. Thus for skewed distributions, range is usually very misleading measure.
- (iv) Since range ignores all the scores except the two extreme scores it cannot be used for making inferences about populations.
- (v) Range varies considerably from sample to sample.

6.4.3 Inter-quartile Range

The inter-quartile range usually denoted by IQR, is a kind of range. It avoids some of the problems associated with R by taking into consideration only the middle half of a distribution. To find IQR:

- (i) Arrange the scores from lowest to highest.
- (ii) Divide the distribution into quartiles and calculate the first, the second, and the third quartiles using the formula discussed in the previous unit.
- (iii) The IQR is defined as the distance between the third quartile Q_3 and the first quartile Q_1 .
Symbolically,

$$\text{IQR} = Q_3 - Q_1$$

Thus IQR extracts the middle half of the cases and then calculates the range. IQR avoids the problem of being based on the most extreme scores by excluding the extremes, but it has all the other disadvantages associated with R.

Check your progress

1. Explain the Measures of Dispersion.
2. What do you mean by range?
3. Define the inter-quartile range.

6.5 Mean Deviation

Average deviation is found by summing the absolute values of the deviations and dividing the sum by number of observations. The formula for average deviation can be written as:

$$MD = \left(\sum |X_i - \bar{X}| \right) / n$$

Where $\bar{X}$ is the arithmetic mean; $X_i - \bar{X}$ is deviation of from $\bar{X}_i$ and $|X_i - \bar{X}|$ is the absolute value of the deviation which is always a positive number. The average deviation tells the distance with which a score will typically deviate from the mean.

Example 2.2: The number of terms that five randomly selected Members of Parliament have served are: 3, 10, 12, 7, 8. Find the average deviation of these scores.

Solution: Make the following table containing the calculation.

Case number	Terms	$X_i - \bar{X}$	$ X_i - \bar{X} $
1	3	3-8=-5	5
2	10	10-8=2	2
3	12	12-8=4	4
4	7	7-8=-1	1
5	8	8-8=0	0
Total	40	0	12

Mean $\bar{X} = 40/5=8$

Sum of deviation from the mean $(X_i - \bar{X}) = 0$

Sum of absolute deviations $|X_i - \bar{X}| = 12$. Therefore

MD= $12/5=2.4$ terms.

For descriptive purpose the average deviations can be an adequate and easily interpretable measure for describing the degree of dispersion. But the mathematical properties of average deviation are

such that it does not meet the needs of advanced mathematics. Therefore, average deviation is a very infrequently used measure of dispersion.

6.6 Variance and Standard Deviation

Instead of taking absolute values of deviations to remove the negative signs and obtain a nonzero sum, another way to get rid of negative sign of deviations is to square them. Square of a negative number is a positive quantity. A statistic called variance uses this approach. To calculate variance, calculate the deviations, square each deviation, add up the squared deviations to obtain sum of squares and divide this sum of squares by the number of deviations. The resulting quantity is called the mean squared deviation (MSD) or the variance of the distribution of scores.

Variance can be two types;

- (i) Sample variance denoted by S^2 , calculated from sample data.
- (ii) Population variance denoted by σ^2 calculated from population data.

In practice S^2 , a statistic, is always known while σ^2 , a parameter is seldom known. Therefore, S^2 is used as an estimate of σ^2 . While dealing with several variables, It proves to be convenient to attach a subscript to s or σ . The subscript indicates the name of variable for which variance is being calculated. Thus S_x^2 is the sample variance of the variable X, S_y^2 is the sample variance of Y, and so on.

6.6.1 Computation of Variance

Variance may be defined as the mean squared deviation of scores around the mean. In the form of a formula, variance is given by:

$$S_x^2 = \sum \frac{(X_i - \bar{X})^2}{n} \quad (\text{for sample data})$$
$$\sigma_x^2 = \sum \frac{(X_i - \mu)^2}{n} \quad (\text{for population data})$$

Where S_x^2 = sample variance of variable X; σ_x^2 is population variance X_i is the value of X variable for ith case; $\bar{X}$ is sample mean; μ is population mean; n is the sample size; and N is population size. The above formula is used only each score has a frequency of 1 and data are ungrouped. If some scores occur more or less frequently than others, a compact ungrouped frequency table may be constructed in which the entries in the column of frequencies are not all the same and they are not all 1's. In such a case the formula for variance is written as:

$$S_x^2 = \sum f_i \frac{(X_i - \bar{X})^2}{n - 1} \quad \text{where } f_i \text{ is the frequency of } X_i$$

$$\sigma_x^2 = \sum f_i \frac{(X_i - \mu)^2}{N} \quad \text{where } f_i \text{ is the frequency of } X_i$$

In calculating the sample variance, the reason for dividing by n-1, instead of n, is to get an unbiased estimate of known population variance. The variability of a sample of scores tends to be less than the variability of the population from which the score are taken. In order to use the sample variance as an unbiased estimate of population variance, a correction factor (n-1) is used in the denominator of the formula for the variance of a sample.

In other words, sample variance almost always underestimates its corresponding population variance and dividing by (n-1), instead of n, tries to compensate for this underestimate. For larger sample sizes (n>100), it makes little difference whether one divides by n or n-1. Significant error can occur if sample is small (n<25). In situations where the interest is merely in describing the variability in the data at hand, only n should be used as a divisor. As a general rule, one can almost use n-1 for sample data. As a descriptive statistic for variability, the variance changes in value as a function of the amount of variability in the data. When all scores are identical the value of variance will be zero. As scores become more dispersed around the mean the value of variance increases. Variance is based on squared deviations and therefore it is always than or equal to zero.

6.6.2 Standard Deviation

Although variance is a very useful measure of variability its value as a descriptive statistic is limited somewhat by the difficulty most people have in thinking about squared deviations. For instance, if you were calculating the variability for income scores (measured in Rs.), the variance

will be expressed in squared units (Rs. Rs or Rs²) and you might obtain a value of variance say 16 Rs. Rs. In the process of squaring the units also get squared. This is what is done when area is reported and calculated square feet square inches etc.

Computing the square root of the variance expresses this variability in terms of the original score values such as Rs. 4, which is easier to interpret and comprehend. This square root of the variance is called the standard deviation (SD), represented by s. The SD is approximately equal to the mean deviation (MD) of scores around the mean, since variance is the mean squared deviation (MSD), standard deviation is root mean squared deviation (RMSD). Because the standard deviation is more readily interpretable than variance, it is used more often to describe data variability.

i.e., Standard deviation = Square root of variance or $s = \sqrt{s^2}$

$$s = \sqrt{s^2} = \sqrt{\left[\sum \frac{(X_i - \bar{X})^2}{n - 1} \right]} = \sqrt{\sum f_i \frac{(X_i - \bar{X})^2}{n - 1}}$$

Similarly, population standard deviation $\sigma = \sqrt{\sigma^2}$.

6.6.3 Effect of Change or Origin and Scale

This S.D. is

$$S_{xi} = \sqrt{\sum \frac{(X_i - \bar{X})^2}{n - 1}}$$

Where $\bar{X} = \frac{1}{n} \sum x_i$

If

$$\delta_i = \frac{x_i - a}{h} \quad \forall i = 1, 2, \dots, n.$$

$$\rightarrow x_i = \delta_i h + a$$

$$\text{then } \bar{X}_{\delta i} = a + \frac{h \sum \delta_i}{n}$$

$$\text{And } (n-1) S^2 = \sum \left[(\delta_i h + a) - \left(a + \frac{h \sum \delta_i}{n} \right) \right]^2$$

$$= \sum \left[\delta_i h - h \left(\frac{\sum \delta_i}{n} \right) \right]^2$$

$$= h^2 \sum (\delta_i - \bar{X}_{\delta i})^2$$

$$S_{xi}^2 = h^2 \frac{\sum (\delta_i - \bar{X}_{\delta i})^2}{n - 1}$$

$$= h^2 S_{\delta i}^2$$

$$S_{xi} = \sqrt{h^2 S_{\delta i}^2} = h S_{\delta i}$$

$$S_{xi} = h S_{\delta i}$$

i.e., standard deviation is independent of change of origin but not independent of change of scale.

Example 2.3: Calculate the mean and S.D. for the following given table of marks distribution of 50 students

Marks	Students (f)	Mid value (x)	$x_i - 25$	$\delta_i = \frac{x_i - 25}{10}$	$f_i \delta_i$	$f_i \delta_i^2$
0-10	2	5	5-25=-20	-2	-4	8
10-20	10	15	15-25=-10	-1	-10	10
20-30	15	25	25-25=0	0	0	0
30-40	14	35	35-25=10	1	14	14
40-50	9	45	45-25=20	2	18	36
	50				18	68

$$\bar{x} \delta_i = a + h \frac{\sum f_i \delta_i}{N} \quad N = \sum f_i$$

$$= 25 + \frac{10 \times 18}{50} = -25 + 3.6 = 28.6$$

$$S_{xi}^2 = h^2 S^2$$

$$\begin{aligned} S_{\delta i}^2 &= \left(\frac{1}{N} \sum f_i \delta_i^2 \right) - \bar{x}_{\delta i}^2 \\ &= \left(\frac{1}{N} \sum f_i \delta_i^2 \right) - \left(\frac{1}{N} \sum f_i \delta_i^2 \right)^2 \\ &= \frac{1}{50} \times 68 - \left(\frac{1}{50} \times 18 \right)^2 \\ &= 1.36 - (0.36)^2 \\ &= 1.36 - 0.1296 = 1.4896 \\ S_{xi}^2 &= 10 \times 1.49 = 14.896 \end{aligned}$$

Example 2.4: For a group of 100 candidates, the mean and standard deviation were found to be 20 and 8 respectively. Later it discovered that no. 23 and 36 misread as 32 and 63. Find the correct mean and S.D. corresponding to the correct numbers.

Solution: Let x be the variable, we have

$$n=100 \quad \bar{x} = 20 \quad S=8$$

Now

$$\begin{aligned} \bar{x} &= \frac{1}{n} \sum x_i \rightarrow \sum x_i = 100 \times 20 = 2000 \\ \text{Corrected } \sum x_i &= 2000 - 2336 + 32 + 63 = 2036 \\ \text{corrected mean} &= \frac{2036}{100} = 20.36 \end{aligned}$$

Similarly,

$$S^2 = \frac{1}{n} \sum x_i + \bar{x}^2$$

$$\sum x_i^2 = n(S^2 + \bar{x}^2)$$

$$\sum x_i^2 = 100 + (64 + 400) = 46400$$

Corrected

$$\sum x_i^2 = 46400 - (23)^2 - (36)^2 + (32)^2 + (63)^2$$

$$= 46400 - 529 - 1296 + 1024 + 3969$$

$$= 49568$$

Corrected $\sum x_i^2 = 49568$.

$$\text{Now corrected } S^2 = \frac{49568}{100} - (20.36)^2$$

$$= 495.68 - 414.5296 = 81.1504.$$

$$\text{corrected } S = 9.0084 \cong 9$$

$$\text{Corrected mean} = 20.36$$

Corrected S=9.

2.6.4 Steps in Computing the Standard Deviation

- (i) Make a frequency distribution table, if not already made, containing two columns, namely, the columns for score values and their frequencies.
- (ii) Calculate the mean score, if not already given:

$$\bar{x} = \frac{\sum f_i X_i}{n} \quad \text{where } n = \sum f_i$$

- (iii) Subtract the mean $\bar{x}$ from each of the scores X_i to calculate deviations. $X_i - \bar{X}_i$ Write these deviations in a separate column, say column 3. Sum all these deviations and see if the sum is zero (excepting the rounding errors.)
- (iv) Square each deviation obtained in step (iii) and write the squared amounts in a separate column say column 4.
- (v) Sum all entries in col-4 to obtain a quantity $\sum f_i(X_i - \bar{X}_i)^2$
- (vi) Take the square root of variance in step (v) to obtain the standard deviation.

Example 2.5: “How accurate are eyewitness reports of accidents?” Social scientists have studied this question in detail. In one experiment, subject viewed a film of an accident in which a car ran a stop sign and hit a parked car. The speed of the car was 31 miles per hour. After viewing the film subjects were asked to estimate the speed of the car. Ten subject gave the following estimates:

15, 40, 32, 18, 35, 20, 37, 35, 28, 40

Calculate the mean and standard deviation for these data. How accurate were the estimates considering the mean score across all subjects? How does the SD help in interpret the mean?

Solution: To calculate SD, it is useful to make the following table

X_i	F_i	$f_i X_i$	X_i^2	$f_i X_i^2$
15	1	15	225	225
40	2	80	1600	3200
32	1	32	1024	1024
18	1	18	324	324
35	2	70	1225	2500
20	1	20	400	400
37	1	37	1369	1369
28	1	28	784	784
Total	10	300		9826

Mean $\bar{X} = \frac{\sum f_i X_i}{n} = \frac{300}{10} = 30$

Sample variance $S_x^2 = [n(\sum f_i X_i^2) - (\sum f_i X_i)^2]/(n)(n - 1)$

$$= \frac{[10(9826) - (300)^2]}{(10)(10 - 1)} = (98260 - 9000)/(10)(9)$$

$$= \frac{8260}{90} = 91.78$$

Standard deviations $= \sqrt{S^2} = \sqrt{(91.78)} = 9.58$

Both variance and standard deviation are based on two important properties of the mean: (i) Sum of the differences of scores from the mean in a distribution equals zero. It is due to this property that the deviations from the mean are squared. (ii) The sum of the squared differences of each value in a distribution from the mean of the distribution yields a minimum value, $\sum f_i (X_i - \bar{X}_i)^2 = \text{minimum}$.

6.6.5 Combined Variance

In n_1 and n_2 be the sizes of two series with respective means $\bar{X}_1, \bar{X}_2$ and respective variances then the standard deviations S_1^2, S_2^2 of the combined series is denoted as S and defined as

$$S^2 = \frac{1}{n_1 + n_2} [n_1(S_1^2 + d_1^2) + n_2(S_2^2 + d_2^2)]$$

Where $d_1 = \bar{x}_1 - \bar{x}, \quad \bar{x}_1 = \frac{1}{n_1} \sum x_{ij}$

$$d_2 = \bar{x}_2 - \bar{x}, \quad \bar{x}_2 = \frac{1}{n_2} \sum x_i,$$

$$\text{and } \bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} \text{ (combined mean)}$$

Another formula is-

$$S^2 = \frac{1}{n_1 + n_2} \left[n_1 S_1^2 + n_2 S_2^2 + \left\{ \frac{n_1 n_2}{n_1 + n_2} (\bar{x}_1 - \bar{x}_2)^2 \right\} \right]$$

Where

$$S_1^2 = \frac{1}{n_1} \sum (x_i - \bar{x}_1)^2$$

$$S_2^2 = \frac{1}{n_2} \sum (x_2 - \bar{x}_2)^2$$

Example 2.6: An analysis of monthly wages paid to the engineers in two companies A and B gives the following results.

	A	B
No. of engineers	1000	2000
Average monthly salary	240.00	275.00
Variance of dist. of salary	41	.80

- (a) Calculate average monthly salary.
 (b) Variance of the distribution of monthly salary of all engineers in A & B taken together.

Solution: $n_1 = 1000$ $n_2 = 2000$

$$\bar{x}_1 = 240 \quad \bar{x}_2 = 275$$

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

$$\bar{x} = \frac{(240 \times 1000) + (275 \times 2000)}{1000 + 2000}$$

$$= \frac{240000 + 550000}{3000}$$

$$= \frac{790000}{3000} = \frac{790}{3}$$

$$= 263.34$$

$$\begin{aligned}
\text{(b)} \quad S^2 &= \frac{1}{n_1+n_2} \left[n_1 S_1^2 + n_2 S_2^2 + \left\{ \frac{n_1 n_2}{n_1+n_2} (\bar{x}_1 - \bar{x}_2)^2 \right\} \right] \\
S^2 &= \frac{1}{3000} \left[41000 + 160000 + \left\{ \frac{4000000}{3000} (240 - 275)^2 \right\} \right] \\
&= \frac{1}{3} \left[201 + \left\{ \frac{4}{3} \times 1225 \right\} \right] \\
&= \frac{1}{3} [201 + 1633.34] \\
&= \frac{1834.34}{3} = 611.45
\end{aligned}$$

Hence $S = 24.73$.

6.6.6 Properties of Standard Deviation

- (i) Standard deviation gives a measure of dispersion relative to the mean.
- (ii) Standard deviation is sensitive to each of the scores in the distribution.
- (iii) Like the mean, standard deviation is stable with regard to sampling fluctuations. This property is one of the main reasons why the standard deviation is used so much often than other measures of dispersion.

6.6.7 Interpretation of Standard Deviation

Standard deviation measures the average dispersion in a data set. It is the average amount by which scores in a distribution deviate from the mean of the distribution. Intuitively large values of standard deviation show that the observations are quite spread out and smaller values indicates that the scores are less dispersed and are clustered around the mean. In an extreme case, for example, a standard deviation of 0 means all of the scores are exactly equal to the mean score and there is no variability among the scores in a distribution.

6.7 Relationship between Measures of Central Tendency and Measures of Dispersion

If $\bar{x}$ be the arithmetic mean, G be the geometric mean, H be the harmonic mean, S^2 be the variance (or S be the standard deviation) and Mol be the mean deviation then for the discrete distribution:

- (i) $G = \bar{x} \left(1 - \frac{1}{2} \cdot \frac{S^2}{\bar{x}^2}\right)$
- (ii) $\bar{x}^2 - G^2 = S^2$
- (iii) $H = \bar{x} \left(1 - \frac{S^2}{\bar{x}^2}\right)$
- (iv) $S^2 > (\text{Ma from mean})^2$.

6.8 Coefficient of Variation (CV)

It is sometimes desirable to compare several groups with respect to their relative homogeneity in instances where the groups have very different means. Therefore, it might be somewhat misleading to compare the absolute magnitudes of the standard deviations. One might expect that with a very large mean one would find a fairly large standard deviation. One might therefore be primarily interested in the size of the standard deviation relative to that of the mean. This suggests that we can obtain a measure of the relative variability by dividing the standard deviation by the mean. The result has been termed the coefficient of variation, denoted by CV. Thus

$$CV = S/\bar{X}$$

Where s is the SD and $\bar{X}$ is the mean.

The coefficient of variation being a ratio requires that one have a ratio level of measurement and not merely interval measurement. You can also realize that one should always report the mean as well as the SD for the data.

The coefficient of variation being a ratio requires that one have a ratio level of measurement and not merely interval measurement. You can also realize that one should always report the mean as well as the SD for the data.

To illustrate the advantages of CV over the SD, suppose a social psychologist is attempting to show that for all practical purpose two groups are equally homogeneous with respect to age. In one group the mean age is 26 with an SD of 3. In the other one, the mean age is 38 with an SD of 5. The coefficient of variation for the two groups are:

$$CV1=3/26=.115, CV2= 5/38=.132$$

The difference between the two coefficients is smaller than the difference between the two SDs. In view of the fact that exact age usually becomes less important in determining interest, abilities and social status as the average age of group members is increased a comparison of the two coefficients of variation in this instance might very well be much less misleading than if the SDs were used.

As another example suppose one is concerned about the dispersions in traffic flows from one weekday to the next at various times of the day. Dispersions in these flows might be misleading in an absolute sense unless standardized by their means so as to allow for differences in the average volumes of traffic at different times of the day.

6.9 Exercises

1. You have just won the state lottery and are now fabulously wealthy. One of the first things you want to do is to find the 'nicest place to live' in all the world. Because you are somewhat eccentric, your only criterion for "nicest place" is climate. Specifically, you want to locate a city where the temperature is exactly 78° . After much search, you find three cities where the average daily temperature is exactly 78° . Based on just this much information. Which of the cities will you choose as your permanent residence? Now suppose you also discovered that the SD and range of the daily temperature were 0.7° and 3° in city A. 10.3° and 30° in city B, 25.8° and 103° in city C. How will this additional information be useful to you in choosing the place you want to live? Can you choose a permanent residence now? Which city would you choose and why?
2. Suppose men and women have about the same distribution of scores on the verbal scholastic aptitude test (SAT), but, on the mathematical part, men have a distinct edge. In 1994, the

men average about 500 on the mathematical SAT, while the women averaged about 460. Both histograms follow the normal curve, with standard deviation of 100.

- (a) Estimate the percentage of men getting over 600 on this test in 1993.
- (b) Estimate the percentage of women getting over 600 on this test in 1993.
- (c) Suppose one of the men who took the mathematical SAT will be picked at random, and you have to guess his test score. You will be given a dollar if you guess it right to within 50 points. What should you guess? What is your chance of winning? Briefly explain your answer.

3. For a normal curve answer the following questions:

- (a) Find the proportion of the area between the mean and a z score of 5.
- (b) What proportion of the area lies to the right of a z scores of .5?
- (c) What proportion of the area lies to the left of a z scores of .5?
- (d) What proportion of the area lies in interval bounded by z scores of -1 and -2?
- (e) Assume the age of state governors in the United States is normally distributed with a mean of 56 and a standard deviation of 8 years. How many governors are between 60 and 70 years of age? (there are 50 governors total).

4. Suppose that for a particular year, on the law school admissions test (LSAT), the mean score for all people taking the test is 500, the standard deviation is 90, and the score are normally distributed. (a) What percentage of people had scores (i) over 600; (ii) less than 300; (iii) between 700 and 750? (b) If a person had a score of 630 on the test, what percentage of people had scores less than his? Greater than his ? (c) If 5000 people took the test that year, how many had a score between 300 and 400?

P-1.6 Suppose that a college entrance examination is given to all entering college students. It is found that the scores are normally distributed with a mean of 450 and a standard deviation of 75. (a) What is the probability that if a student were selected at random, his score in the

test would be (i) greater than 450; (ii) less than 550; (iii) between 350 and 400? (b) If the z core of a student on this test were- 1.5, what was his original score?

6.10 Summary

Various measures of dispersion have been defined and formulas for their calculation are given in this unit. Once data have been represented by a measure of central tendency, one may like to know the scatter of the given data around this measure of central tendency. The various measures of dispersion are range, quartile deviation, mean deviation, standard deviation and variance. Coefficient of variation for consistency of data or frequency distribution is defined as the ratio of standard deviation to arithmetic mean. It has no unit.

6.11 Further Readings

1. Goon A.N., Gupta M.K. & Das Gupta B (1987) *Fundamentals of Statistics Vol. I* The World Press Pvt. Ltd., Kolkata.
2. Yule G.U. and Kendall, M.G.: *An Introduction to the Theory of Statistics*, Charles Griffin and Co. Limited.



Master of Science PGMM -108N Mathematical Statistics

U. P. Rajarshi Tandon
Open University

Block

3 Moments, Skewness and Kurtosis

Unit- 7	Moments and Raw Moments	151-158
Unit- 8	Central Moments	159-178
Unit- 9	Skewness and Kurtosis	179-194

Block-3

Moments, Skewness and Kurtosis

The **Block - 3 – Moments, Skewness and Kurtosis** deals with moments, skewness and kurtosis. It consists of two units. The *first unit* of this block defines various moments of the frequency distribution and give the interrelationship between them. The *second unit* of this block deal with Central moments expressed in terms of raw moments, Raw Moments expressed in terms of central moments, Effects of change of Origin and Scale on Central Moments, Charlier's Check, and Shephard's Corrections for Moments.

The *second unit* of this block explains the significance of asymmetrical data and describes various measures of skewness i.e. lack of symmetry of data. It also gives the measures of kurtosis and explains the peakedness of the frequency curve near the highest frequency.

At the end of every block/unit the summary, self-assessment questions and further readings are given.

Unit-7: Moments, Raw Moments and Central Moments

Structure

7.1 Introduction

7.2 Objectives

7.3 Moments (Definition)

7.3.1 Raw Moments for Ungrouped data

7.3.2 Raw Moments for grouped data

7.4 Some Solved Examples

7.5 Summary

7.6 Exercises

7.7 Further Readings

7.1 Introduction

Measures of central tendency and variability (dispersion) enable us to know some important characteristic of the data and help us to compare two or more series. It can be illustrated that two different distributions may have the same mean and/or variance, still they may have different pattern of the distribution. Two other characteristics of the distribution are known as symmetry and peakedness of the curve. These may be defined in terms of the moments of the distribution.

The word moment is derived from statics, where moment about a point is equal to force multiplied by the perpendicular distance. The mean and variance of the distribution are the first moment about origin and second moment about mean or second central moment of the distribution.

7.2 Objectives

After going through this unit, you shall be able to

- Compute raw moments including mean of the given data/frequency distribution
- Compute the raw moments for ungrouped and grouped data

7.3 Moments (Definition)

Suppose we have n values of a variables X as $X_1, X_2, \dots, X_n$. The possible measures of central tendency and dispersion of variable x are mean and variance defined by expression:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \text{ and } \sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

$\bar{X}$ is the first moment of $\underline{X}$ about the origin.

7.3.1 Raw Moments for Ungrouped Data

If $X_1, X_2, \dots, X_n$ are n values of the variable x, the r^{th} raw moment of x about any point A is defined as-

$$m'_r = \frac{1}{n} \sum_{i=1}^n (x_i - A)^r ; r = 0, 1, 2 \dots \dots \dots (1.1)$$

So, (always)

$$\left. \begin{aligned} m'_0 &= \frac{1}{n} \sum (x_i - A)^0 = 1 \\ m'_1 &= \frac{1}{n} \sum (x_i - A)^1 = (\bar{x} - A) \\ m'_2 &= \frac{1}{n} \sum (x_i - A)^2 \dots \dots \dots \\ m'_3 &= \frac{1}{n} \sum (x_i - A)^3 \\ m'_4 &= \frac{1}{n} \sum (x_i - A)^4 \end{aligned} \right\} (1.2)$$

In particular, if the r^{th} raw data about origin, i.e. for $A = 0$ is

$$m'_r = \frac{1}{n} \sum x_i^r$$

so that,

$$m'_0 = \frac{1}{n} \sum x_i^0 = 1 \text{ (always)}$$

$$m'_1 = \frac{1}{n} \sum x_i = \text{mean of the distribution}$$

$$m'_2 = \frac{1}{n} \sum X_i^2$$

$$m'_3 = \frac{1}{n} \sum X_i^3$$

$$m'_4 = \frac{1}{n} \sum X_i^4.$$

Check your progress

1. What do you mean by moments?

7.3.2 Raw Moments for the grouped Data

If the given values are in the form of a frequency distribution,

Table 1.1

Value of X_i (X)	$X_1, X_2, \dots, X_i, \dots, X_n$
Frequency	$f_1, f_2, \dots, f_i, \dots, f_n$

The formula for moments about the point A takes the form

$$m'_r = \frac{1}{n} \sum_{i=1}^n (x_i - A)^r; \quad r = 0, 1, 2, \dots \dots (1.3)$$

Table 1.2

Class Interval	Mid-point of the class	frequency
	X_1	f_1
	X_2	f_2

	‘	‘
	‘	‘
	‘	‘
	X _i	f _i
	‘	‘
	‘	‘
	‘	‘
	X _n	f _n
Total	----	N=∑ f _i

We shall write it as “ X_i/f_i (I=1,2,.....n) distribution”

Where X_i is class mark of the ith class, or its value of the variable X (Table 1.1), f_i is its frequency and N= ∑ f_i is total frequency. (number of observations)

$$m'_0 = \frac{1}{n} \sum f_i x_i^0 = \frac{1}{n} \sum f_i = 1 \text{ (always)}$$

$$m'_1 = \frac{1}{n} \sum f_i x_i = \text{mean of the distribution} = \bar{x}$$

$$m'_2 = \frac{1}{n} \sum f_i X_i^2$$

$$m'_3 = \frac{1}{n} \sum f_i X_i^3$$

$$m'_4 = \frac{1}{n} \sum f_i X_i^4.$$

7.4 Some Solved Examples

Example 1.1: Find first four raw moments about x= 5 if the values of variables are 2,3,6,8 and 11.

Solution: Here $m'_0 = 1$

$$m'_1 = \frac{1}{N} \sum (x_i - A) = (\bar{x} - A)$$

$$\bar{x} = \frac{2 + 3 + 6 + 8 + 11}{5} = \frac{30}{5} = 6$$

$$m'_1 = 6 - 5 = 1$$

$$m'_2 = \frac{1}{N} \sum_{i=1}^n (x_i - A)^2$$

$$= \frac{(2-5)^2 + (3-5)^2 + (6-5)^2 + (8-5)^2 + (11-5)^2}{5}$$

$$= \frac{9 + 4 + 1 + 9 + 36}{5} = \frac{59}{5} = 11.8$$

Similarly, $m'_3 = \frac{(2-5)^3 + (3-5)^3 + (6-5)^3 + (8-5)^3 + (11-5)^3}{5}$

$$= \frac{-27 - 8 + 1 + 27 + 216}{5} = \frac{209}{5} = 41.8$$

And $m'_4 = \frac{(2-5)^4 + (3-5)^4 + (6-5)^4 + (8-5)^4 + (11-5)^4}{5} = 295$

Hence, the first four moments are 1, 11.8, 41.8 and 295.

Example 1.2: The number of suits sold daily by a women's boutique on the past six days has been given in the following frequency table.

Value (x)	Frequency (f)
3	2
4	1
5	3

Obtain the first four raw moments about origin.

Solution: r^{th} raw moments about origin is given by

$$m'_{r_i} = \frac{1}{N} \sum f_i x_i^r ; \quad r = 0, 1, 2, \dots$$

$$m'_0 = 1$$

$$\begin{aligned} m'_{r_i} &= \frac{1}{N} \sum f_i x_i^r = \frac{(2 \times 3) + (1 \times 4) + (3 \times 5)}{6} \\ &= \frac{25}{6} = 4.17 \end{aligned}$$

Similarly

$$m'_2 = \frac{2 \times 3^2 + 1 \times 4^2 + 3 \times 5^2}{6} = \frac{109}{6} = 18.17$$

$$m'_3 = \frac{2 \times 3^3 + 1 \times 4^3 + 3 \times 5^3}{6} = 82.17$$

$$m'_4 = \frac{2 \times 3^4 + 1 \times 4^4 + 3 \times 5^4}{6} = 382.17$$

Hence the first four raw moments are 4.17, 18.18, 82.17 and 382.17 suits respectively.

7.5 Summary

In this unit the properties of moments are discussed. In probability theory and statistics, the raw moments of a random variable are a set of quantities that describe the shape of its probability distribution. Raw moments are used to calculate central moments, which provide more information about the distribution by centering the moments around the mean.

7.6 Exercises

Q.1. Define raw moments.

- Q.2. The first three moments of a distribution about the value 2 of the variable are 1, 16 and -40. Show that mean is 3, the variance is 15 and $\mu_3 = -86$.
- Q.3. The first three moments of a distribution about the value $x=7$ are 3, 10 and 15 respectively. Obtain mean, variance and m_3 . (Ans. 10, 1, -2.1).
- Q.4. The first four raw moments of a distribution about $x=4$ is 1, 4, 10, 45. Show that the mean is 5 and the variance is 3 and m_3 and m_4 are 0 and 26 respectively.

7.7 Further Readings

1. Goon A.M., Gupta M.K. & Das Gupta B (1987) Fundamentals of Statistics Vol. I The World Press Pvt. Ltd., Kolkata.
2. Yule G.U. and Kerdall, M.G.: An Introduction to the Theory of Statistics, Charles Griffin and Company Ltd.

Unit-8 : Central Moments

Structure

8.1 Introduction

8.2 Objectives

8.3 Central Moments

8.4 Factorial Moments

8.5 Inter-relationship between various moments

8.5.1 Central moments expressed in terms of raw moments.

8.5.2 Raw Moments expressed in terms of central moments.

8.6 Effects of change of Origin and Scale on Central Moments

8.7 Charlier's Check

8.8 Shephard's Corrections for Moments

8.9 Some Solved Examples

8.10 Exercises

8.11 Answer and Suggestion

8.12 Summary

8.13 Further Readings

8.1 Introduction

Measures of central tendency and variability, also known as dispersion, help us understand important characteristics of data and enable comparisons between two or more datasets. It's possible for two different distributions to have the same mean and/or variance, yet exhibit different patterns. Symmetry and peakedness of the curve are two additional characteristics of a distribution, which can be defined in terms of its central moments.

The term "moment" is derived from statics, where a moment about a point equals the force multiplied by the perpendicular distance. The mean and variance of a distribution correspond to the first moment about the origin and the second moment about the mean (or the second central moment) of the distribution.

8.2 Objectives

After going through this unit, you shall be able to

- Compute central moments including variance of the given data/frequency distribution
- Compute factorial moments of the given data/frequency distribution.
- Use the interrelationship between these moments to obtain one for; the other known moments.
- Apply Charlier's check and Shephard's correction for moments.

8.3 Central Moments

If the arbitrary origin of moments of variable X is taken as arithmetic mean *i.e.*, $A = \bar{x}$ the moments are called central moments.

For ungrouped data $X_1, X_2, \dots, X_n$ the r^{th} central moment of variable X is given by

$$m'_r = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^r; \quad r = 0, 1, 2, \dots \quad (1.5)$$

If the given values are classified into a frequency distribution, X_i/f_i ($i=1, 2, \dots, n$), the r^{th} central moment is given by

$$m_r = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^r; \quad r = 0, 1, 2, \dots \quad (1.6)$$

X_i being the mid-value of i^{th} or x^{th} value of the variable (as the case may be) class and f_i its frequency

Evidently, we have

$$m_0 = 1$$

and

$$m_1 = 0 \text{ (always)}$$

The second central moment of variable X is variance of the distribution *i.e.*

$$v(x) = m_2 = \left\{ \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2; \quad (\text{for ungrouped data}) \right.$$

$$\left. \frac{1}{n} \sum_{i=1}^n f_i (x_i - \bar{x})^2; \quad (\text{for ungrouped data}) \dots \dots \dots (1.8) \right.$$

Third and fourth central moments for ungrouped and grouped data are-

$$\left\{ \begin{array}{l} m_3 = \left\{ \frac{1}{n} \sum_{i=1}^n (x_1 - \bar{x})^3; \quad (\text{for ungrouped data}) \right. \\ \\ m_3 = \left\{ \frac{1}{n} \sum_{i=1}^n f_i (x_1 - \bar{x})^3; \quad (\text{for grouped data}) \dots \dots \dots (1.9) \right. \end{array} \right.$$

$$\left\{ \begin{array}{l} m_4 = \left\{ \frac{1}{n} \sum_{i=1}^n (x_1 - \bar{x})^4; \quad (\text{for ungrouped data}) \right. \\ \\ m_4 = \left\{ \frac{1}{n} \sum_{i=1}^n f_i (x_1 - \bar{x})^4; \quad (\text{for grouped data}) \dots \dots \dots 1.10 \right. \end{array} \right.$$

8.4 Factorial Moments

The r^{th} factorial moment about the origin of the distribution X_i/f_i ($i=1,2,\dots,n$), is defined as follows:

$$\mu_{(r)} = \frac{1}{n} \sum_{i=1}^n f_i x_1^{(r)}$$

$$\text{where } x^{(r)} = x(x-1)(x-2) \dots (x-r+1) \text{ and } \frac{1}{n} \sum_{i=1}^n f_i \dots \dots (1.11)$$

Similarly the factorial moment of order r about any point $x=A$ is given as

$$\mu_{(r)} = \frac{1}{n} \sum_{i=1}^n f_i (x_1 - A)^{(r)}$$

$$\text{where } (x-A)^{(r)} = (x-A)(x-A-1)(x-A-2) \dots (x-A-r+1) \dots \dots (1.12)$$

Check your progress

1. Explain the central moments.
2. What do you mean by factorial moments?

8.5 Inter-Relationship between Various Moments

The moments of a probability distribution are related to each other in several ways, and these relationships can provide insights into the properties of the distribution.

8.5.1 Central moments expressed in terms of raw moments

We have first r raw moments $m'_1, m'_2, \dots, m'_r$ about an arbitrary origin A . The r^{th} order central moment can be obtained by using simple algebra.

We have $(x_i - \bar{x})^r = \{(x_i - A) - (\bar{x} - A)\}^r$

$$\begin{aligned} &= (x - A)^r - r_{c_1}(x_i - A)\}^{r-1}(\bar{x} - A) + r_{c_2}(x_1 - A)\}^{r-2}(\bar{x} - A)^2 \\ &\quad + (-1)^r r_{c_r}(\bar{x} - A)^r \dots \dots \dots (1.13) \end{aligned}$$

Since $m'_1 = \frac{1}{n} \sum f_i(x_1 - A)^1 = (\bar{x} - A)$

The r^{th} central moment can be written as

$$\begin{aligned} m'_1 &= \frac{1}{n} \sum f_i(x_1 - \bar{x})^r \\ &= \frac{1}{n} \sum f_i(x_1 - A)^1 - r_{c_1} \frac{1}{n} \sum f_i(x_1 - A)^{r-1} m'_1 \\ &\quad + r_{c_r} \frac{1}{n} \sum f_i(x_1 - A)^{r-2} m_1^2 \dots \dots + (-1)^r m_1^r \\ \therefore m_r &= m'_r - r_{c_1} m'_{r-1} m'_1 + r_{c_2} m'_{r-2} m_1^2 \dots \dots + (-1)^r m_1^r \dots \dots (1.14) \end{aligned}$$

It is easily seen that (1.14) holds for moments obtained from ungrouped data as well.

Putting r= 1,2,3, and 4 in (1.14), we get some particular cases for A=0, as

$$\begin{aligned}
 m_1 &= m'_1 - m'_1 = 0 \\
 m_2 &= m'_2 - 2m'_1 m'_1 + m_1'^2 \\
 &= m'_2 - m_1'^2 \\
 m_3 &= m'_3 - 3m'_2 m_1 + 3m'_1 m_1^2 - m_1^3 \\
 &= m'_3 - 3m'_2 m_1 + 3m'_1 + 2m_1^3 \\
 m_4 &= m'_4 - 4m'_3 m'_1 + 6m'_2 m_1^2 - 4m'_1 m_1^3 + m_1^4 \\
 &= m'_4 - 4m'_3 m'_1 + 6m'_2 m_1^2 - 4m_1^4 \dots \dots (1.15)
 \end{aligned}$$

If most of the practical problems, it is sufficient to calculate $\bar{x}$, m_2 , m_3 and m_4 using calculators. These computations are greatly facilitated by first compiling moments about a suitable chosen origin A or origin 'O'. We first calculate

$$N = \sum f_i$$

$$\begin{array}{ll}
 \sum f_i x_i = & \sum f_i u_i = \\
 \sum f_i x_i^2 = & \sum f_i u_i^2 = \\
 \sum f_i x_i^3 = \text{or} & \sum f_i u_i^3 = \\
 \sum f_i x_i^4 = & \sum f_i u_i^4 =
 \end{array}$$

Where, and then use $\bar{x}$, m_2 , m_3 , m_4 (1.15) to compute using relations (1.14) and (1.15). For the grouped data $u_i = \frac{x_i - A}{h} \dots \dots \dots (1.16)$

h is the common class interval, and A is a suitably chosen origin or reference point theoretically, A can be any point, but it is so chosen that the computation work may be reduced. For grouped data, A is taken at a mid-point near the central classes.

8.5.2 Raw Moments expressed in terms of central moments

Just as central moments can be expressed in terms of moments about an arbitrary origin A, so a moment about an arbitrary origin is expressible in terms of central moments. From (1.2)

$$\begin{aligned}m'_i &= \bar{x} - A, \text{ and} \\m'_r &= \frac{1}{n} \sum f_i(x_1 - A)^r \\&= \frac{1}{N} \sum_i f_i(x_1 - \bar{x} + \bar{x} - A)^r \\&= \frac{1}{N} \sum_i f_i\{(x_1 - \bar{x}) + m_1\}^r \\&= \frac{1}{N} \sum_i f_i(x_1 - \bar{x}) + r_{c_1}(x_i - \bar{x})^{r-1}m'_1 + r_{c_2}(x_i - \bar{x})^{r-2}m_1^2 + m_1^r \\&= \frac{1}{N} \sum_i f_i(x_1 - \bar{x})^r + r_{c_1} \frac{1}{N} \sum_i f_i(x_1 - \bar{x})^{r-1}m'_1 + r_{c_2} \sum_i f_i(x_1 - \bar{x})^{r-2}m_1^2 + m_1^r \\&= m_r + r_{c_1}m_{r-1}m_1^1 + r_{c_2}m_{r-2}m_1^2 + \cdots \cdots \cdots m_1^r \quad (1.16)\end{aligned}$$

In particular,

$$u_2 = u_2 + u_1^2$$

$$u_3 = u_3 + 3u_2u_1 + u_1^3$$

$$u_4 = u_4 + 4u_3u_1 + 6u_2u_1^2 + u_1^4$$

These formula help us to obtain the moments about any point A , if the central moments are known.

8.6 Effect of change of origin and scale on central moments

If we change the origin of x on some point A and scale by h. The new variable u is defined $u = \frac{x-A}{h}$ as so that $x=A+ hu$, $\bar{x}=A + h\bar{u}$ and

$$\begin{aligned}
m_r &= \frac{1}{n} \sum f_i (x_1 - \bar{x})^r \\
&= \frac{1}{n} \sum f_i (hu_i - h\bar{u})^r \\
&= h^r \frac{1}{N} \sum f_i (u_i - \bar{u})^r \\
&= h^r m_r(u) \dots \dots \dots (1.17) \\
\text{where } m_r^{(4)} &= \frac{1}{N} \sum f_i (u_i - \bar{u})^r
\end{aligned}$$

Thus, r^{th} central moment of variable X is h^r times r^{th} central moments of variable U. So we conclude that central moment is unaffected by change of origin but it is affected by change of scale.

8.7 Charlier's Checks

Charlier's checks are often used as a ready check against some possible mistake in the calculation of first four moments.

For this we first compute $\sum f_i x_i$, $\sum f_i x_i^2$, $\sum f_i x_i^3$, $\sum f_i x_i^4$, etc. and verify the calculations by the following identity.

$$\sum_i f_i (x_i + 1)^2 = \sum_i f_i x_i^4 + 4 \sum_i f_i x_i^3 + 6 \sum_i f_i x_i^2 + 4 \sum_i f_i x_i + N \dots \dots (1.18)$$

8.8 Shephard's Correction for moments

In computing moments for data grouped into class-intervals by means of the formulae

$$m'_r = \frac{1}{n} \sum f_i (x_1 - A)^r \quad \text{and} \quad m_r = \frac{1}{n} \sum f_i (x_1 - \bar{x})^r \quad \dots (1.19)$$

In the computation of mean, variance and various moments from a grouped data into class intervals, we have taken mid-point or class mark as the representative of that class. Here, the assumption is that the midpoint of the classes are reasonable approximations to the mean of the observation which are class. This approximation holds for good for the distribution which are symmetrical, moderately skewed and have classes having small classes intervals. It does not hold for all distributions.

We have the assumption that the observation falling in a class (e.g. the f_i values falling in the i th class) were all equal to the class mark or mid-point, although the observation may be really unequal. The assumption naturally introduces some error, which are called the errors due to grouping. To correct for these grouping errors, the computed values of the moments have to be suitable 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 and for central moments of the first orders 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
$$m_2(\text{corrected}) = m_2 - \frac{c^2}{12}$$

$$m_3(\text{corrected}) = m_3$$

$$m_4(\text{corrected}) = m_4 - \frac{c^2}{2} m_2 + \frac{2}{240} c^4$$

Where c is the width of each class-interval and m_r 's are the uncorrected r^{th} moments for $r=1,2,3,\dots$

The Sheppard's corrections is applicable only when –

- (i) Class-width are equal
- (ii) The distributions are symmetrical or moderately skewed.
- (iii) N is sufficiently large,
- (iv) The frequency curve is not J or U shaped
- (v) It is necessary that the observation should relate to a continuous variable.

8.9 Some Solved Examples

Example.1: The following frequency table gives the values obtained in 15 throws of a die.

Value (x)	Frequency (f)
1	3
2	2
3	3
4	4
5	2
6	1

Find first four central moments.

x_i	f_i	$x_i f_i$	$(x_i - \bar{x})$	$f_i(x_i - \bar{x})^2$	$\sum f_i(x_i - \bar{x})^3$	$\sum f_i(x_i - \bar{x})^4$
1	3	3	-2.2	14.52	-31.94	70.27
2	2	4	-0.2	2.88	-3.46	4.15
3	3	9	-0.2	0.12	-0.02	0.00
4	4	16	0.8	2.56	2.05	1.64
5	2	10	1.8	6.48	11.66	20.99
6	1	6	2.8	7.84	21.95	61.46
Total	N=15	48		34.40	0.24	158.51

$$\therefore \bar{x} = \sum_{i=1}^n \frac{f_i x_i}{N} = \frac{48}{15} = 3.2$$

$$m_1 = 0$$

$$m_2 = \sum_{i=1}^n \frac{f_i(x_i - \bar{x})^2}{N} = \frac{34.40}{15} = 2.293$$

$$m_3 = \sum_{i=1}^n \frac{f_i(x_i - \bar{x})^3}{N} = \frac{0.24}{15} = 0.016$$

$$m_4 = \sum_{i=1}^n \frac{f_i(x_i - \bar{x})^4}{N} = \frac{158.51}{15} = 10.567$$

Answer: First four central moments are $m_1=0$, $m_2= 2.293$, $m_3= 0.016$, $m_4= 10.567$.

Example.2: The first four raw moments of a distribution about the value 5 of variable are 2, 20, 40 and 50. Obtain mean and first four central moments.

Solution: In the above example $A = 5, m'_1 = 2, m'_2 = 20, m'_3 = 40, \text{ and } m'_4 = 50$,

$$m'_1 = \bar{x} - A \rightarrow m'_1 + A = 2 + 5 = 7 = \bar{x}$$

Using relations in equations (2.9) we have

$$m_2 = m'_2 - m_1^{-2} = 20 - 2^2 = 16$$

$$m_3 = m'_3 - 3m_1^2 + 3m_1^3$$

$$= 40 - 3 \times 20 \times 2 + 2^3 = -64$$

$$m_4 = m'_4 - 4m'_3m'_1 + 6m'_2m_1^2 - 3m_1^4$$

$$= 50 - 4 \times 40 \times 2 + 6 \times 20 \times 2^2 - 3 \times 2^4 = 162$$

Answer : $\bar{x} = 7, m_2 = 16, m_3 = -64$ and $m_4 = 162$

Example.3: Calculate the first four central moments of the following distribution.

x:	2	3	4	5	6	7	8	9
f:	1	8	28	56	70	28	8	1

We should shift the origin at a point say A to reduce the calculation. Let us choose A =5, (near the middle of the table). We may have taken $A = \frac{5+6}{2} = 5.5$ but it will make the calculation more cumbersome.

Solution: This is a symmetrical distribution with A = 5 (near the middle of the table), and defining $u=x-A$, we get

$$\sum f_i \mu_i = 0 \text{ and } \sum f_i \mu_i^3 = 0$$

It has simplified the calculation = 0

Calculations

x	f	u=x-5	fu	fu ²	fu ³	fu ⁴
1	1	-4	-4	16	-64	256
2	8	-3	-24	72	-216	648
3	28	-2	-56	112	-224	448
4	56	-1	-56	56	-56	56
5	70	0	0	0	0	0
6	56	1	56	56	56	56
7	28	2	56	112	224	448
8	8	3	24	72	216	648
9	1	4	4	16	64	256
Total	256		0	512	0	2816

$$m'_1 = \sum_i \frac{f_i u_i}{N} = 0; \quad m'_2 = \frac{\sum f_i u_i^2}{N} = \frac{512}{256} = 2$$

$$m'_3 = \frac{\sum f_i u_i^3}{N} = 0; \quad m'_4 = \frac{\sum f_i u_i^4}{N} = \frac{2816}{256} = 11$$

$$m_1 = 0$$

$$m_2 = m'_2 - m_1^2 = 2$$

$$m_3 = m'_3 - 3m'_2m'_1 + 3m_1^2 = 0$$

$$m_4 = m'_4 - 4m'_3m'_1 + 6m'_2m_1^2 - 3m_1^4 = 11$$

Since central moments is unaffected by change of origin the first four central moments of variable X are for given distribution is

$$M'_1(x) = A + M_1^{(u)} = 5, m_2 = 2, m_3 = 0, m_4 = 11$$

You may try the following problems.

8.10 Summary

Central moments are useful because they provide information about the spread and shape of the distribution that is independent of the location (mean) of the distribution. In probability theory and statistics, the central moments of a random variable are a set of quantities that describe the shape of its probability distribution, but with respect to the mean of the distribution. The second central moment is the variance, which measures the spread of the distribution. The square root of the variance is the standard deviation, which is a widely used measure of dispersion.

8.11 Exercises

- P-1.1 Define raw and central moments of a frequency distribution. Obtain the relationship between the central moment of order r in terms of the raw moments. What is Sheppard Corrections to the central moments?
- P-1.2 Define moments. Express central moments in terms of raw moment and vice-versa.
- P- 1.3 The first three moments of a distribution about the value 2 of the variable are 1, 16 and -40. Show that mean is 3, the variance is 15 and $\mu_3 = -86$.
- P-1.4 Express first four central moments in terms of raw moments. What is the effect of change of origin and scale on central moments?

P-1.5 The central moments of a frequency distribution of incomes of a labour class are $m_2=100 \text{ Rs}^2$, $m_3=-4\text{Rs}^2$ and $m_4= -624\text{Rs}^4$.

(a) If the income of each labour is increased by 10 Rs. What will be the moments?

(b) If the income of each labour is doubled. What will be new moments?

P-1.6 What is Sheppard's Correction? What will be the corrections for the first four raw moments and first four central moments?

P-1.7 Find the second, third and fourth central moments for the frequency distribution given below:

Class Interval	Frequency
110-115	05
115-120	15
120-125	20
125-130	35
130-135	10
135-140	10
140-145	05

Also apply Sheppard Correction for moments.

P-1.8 Frequency distribution of scores in mathematics of 50 students are given below:

Score	50-60	60-70	70-80	80-90	90-100
Frequency	1	0	0	1	1
Score	100-110	110-120	120-130	130-140	140-150
Frequency	2	1	0	4	4
Score	150-160	160-170	170-180	180-190	190-200
Frequency	2	5	10	11	4
Score	200-210	210-220	220-230		
Frequency	1	1	2		

Compute first four central moments. Obtain corrected central moments after applying the Sheppard corrections.

(Answer: $m_2=1,176$; $m_3 = -41,160$; $m_4= 57,45,600$; corrected moments are $m_2= 1167.67$; $m_3=-41160$ and $m_4 = 5687091$).

Answer and Suggestions

P- 1.5 (a) Not change m , m_2 , m_3 , m_4

(b) Here $h= 2$, therefore, $m_2=400 R_s^2$, $m_3=-32R_s^3$ and $m_4= =9984R_s^4$.

P-1.7 Solution is shown below:

Class interval	Mid point x	$\mu_i = \frac{x_i - 127.5}{5}$	Frequency f_i	$f_i \mu_i$	$f_i \mu_i^2$	$f_i \mu_i^3$	$f_i \mu_i^4$
110-115	112.5	-3	05	-15	+45	-135	405
115-120	117.5	-2	15	-30	+65	-120	240
120-125	122.5	-1	20	-20	+20	-20	+20
125-130	127.5	0	35	0	0	0	0
130-135	132.5	1	10	10	10	-1	10
135-140	137.5	2	10	20	40	80	160
140-145	142.5	3	05	15	45	135	405
Total			N=100	-20	220	-50	1240

For distribution for μ

Raw moments

$$\mu_i^{(\mu)} = \frac{-20}{100} = -0.2,$$

$$\mu_2^{(\mu)} = \frac{220}{100} = +2.20$$

$$\mu_i^{(\mu)} = \frac{-50}{100} = -0.5,$$

$$\mu_2^{(\mu)} = \frac{1240}{100} = +12.40$$

Central moments

$$\mu_2(\mu) = \mu_i^2(\mu) = 2.20 - (-0.2)^2 = 2.20 - 0.04 = 2.16$$

$$\begin{aligned}\mu_3(\mu) &= \mu_3(\mu) - 3\mu_2(\mu)\mu_1(\mu) + 2\mu_1(\mu) = -0.5 - 3 \times 2.20 \times (-0.2) \times +(-0.2)^3 \\ &= -0.5 + 1.32 - 0.016 = 0.804\end{aligned}$$

$$\begin{aligned}\text{Similarly, } \mu_4(\mu) &= 12.40 - 4 \times (-0.5)(-0.2) + 6(2.20)(-0.2)^2 - 3(-0.2)^4 \\ &= 12.40 - 0.4 + 0.528 - 0.0048 = 12.5232\end{aligned}$$

Hence, for the given distribution, with $h=5$.

$$\mu_2 = 5^2 \times 2.16 = 54.0$$

$$\mu_3 = 5^3 \times 0.804 = 100.5$$

$$\mu_4 = 5^4 \times 12.5232 = 7827.0$$

$$\text{Sheppard correction. } \overline{\mu_2} = m_2 - \frac{h^2}{12} = 54 - \frac{5^2}{12} = 54.02 = 51.9167$$

$$\overline{\mu_3} = m_3 = 100.50$$

$$\overline{\mu_4} = m_4 - \frac{h^2}{2}m_2 + \frac{2}{240}h^2 = 7827 - \frac{25}{2} \times 54 \times \frac{2}{240} \times 5^4 = 7157.2083$$

8.12 Further Readings

1. Goon A.M., Gupta M.K. & Das Gupta B (1987) Fundamentals of Statistics Vol. I The World Press Pvt. Ltd., Kolkata.
2. Yule G.U. and Kerdall, M.G.: An Introduction to the Theory of Statistics, Charles Griffin and Company Ltd.

Unit-9: Skewness and Kurtosis

Structure

9.1 Introduction

9.2 Objectives

9.3 Skewness and its measures

9.4 Measures of skewness

9.4.1 Pearsons's coefficient

9.4.2 Bowley's coefficient

9.4.3 β and γ

9.4.4 Another measures based upon moments.

9.5 Kurtosis

9.5.1 Measures of Kurtosis

9.6 Some solved examples

9.7 Exercises

9.8 Answer and Suggestions

9.9 Summary

9.10 Further Readings

9.1 Introduction

To have an idea about the shape of the frequency or probability curve we study the skewness and kurtosis of the distribution. A distribution is said to be symmetrical if the frequencies (probabilities) are on either side of the central value. It implies that both the right and left tails of the tails of the curve are exactly equal in shape and length. If a distribution is not symmetrical then it is called asymmetric or skewed in the direction of the extreme values, i.e., on the right – or on the left. Since extreme values give longer tail in its direction therefore, the distribution having longer right tail is called right skewed or positively skewed distribution. The left implies longer left tail. Thus, a measure of skewness indicates the extent as well as direction of skewness of the distribution.

Karl Pearson called a normal curve as mesokurtic, it has a hump at the middle. Pearson defined Kurtosis at the convexity of the curve β_2 and γ_2 used and as its measure. The measure of kurtosis gives an idea whether the centre of the distribution is assuming flatness or peakedness similar to the hump of the normal probability curve or not.

The measures of skewness are very useful in biological, chemical and physical laboratory works. They are used in economic social statistics and medical statistics to study the behavior of the data.

9.2 Objectives

After going through this unit, you shall be able to:

- Differentiate between the behavior of symmetrical data and right or left skewed data.
- Obtain the measures of skewness and kurtosis and interpret them.

9.3 Skewness and Its measures

Definition: By skewness of a frequency distribution we mean the degree of its departure from symmetry. The frequency distribution of a discrete variable x is called symmetrical about the value x_0 if the frequency of $x_0 - h$ is the same as the frequency $x_0 + h$ of whatever h may be.

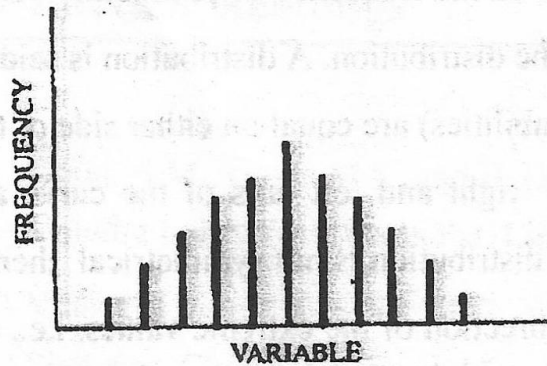


Fig. 2.1a A symmetrical distribution (discrete variable).

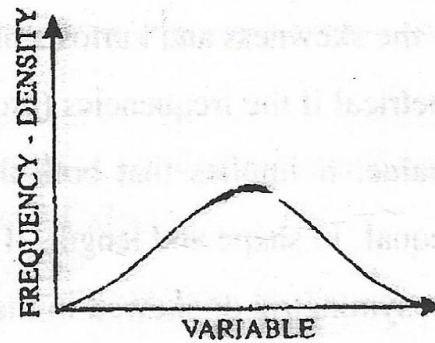


Fig. 2.1b A symmetrical distribution (continuous variable).

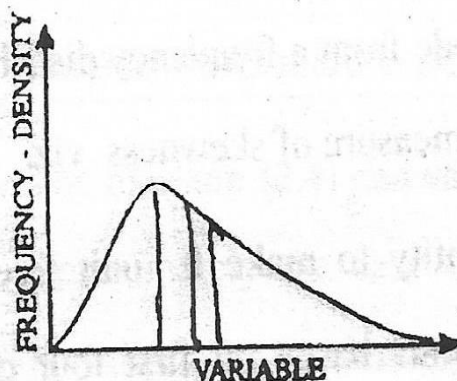


Fig. 2.2a A positively skew distribution.

Mode < Median < Mean

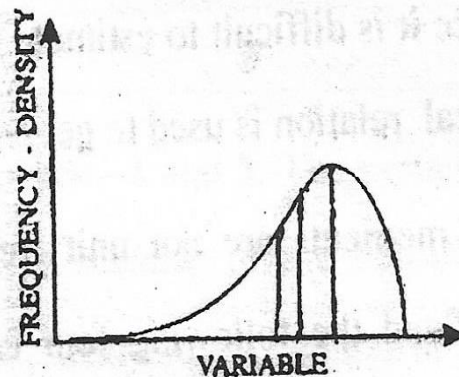


Fig. 2.2b A negatively skew distribution.

Mode < Median < Mean

In the case of a continuous variable, the term 'symmetry' should be used in relation to its frequency curve. The frequency curve of a continuous variable is said to be symmetrical about x_0 if the

frequency density at x_0-h is the same as the frequency- density x_0+h at whatever h may be. Figure 2.1a and 2.1b show two symmetrical distributions.

A distribution which is not symmetrical is called asymmetrical or skew. This skewness is said to be positive if the longer tail of the distribution is towards the higher values of the variable (Fig. 2.2a), negative if the longer tails is towards the lower values of the variable (Fig. 2.2b)

An important point to be noted in this connection is that all odd-order central moments are zero for a symmetrical distribution, positive for a positively skew distribution and negative for a negatively skew distribution. Any such moment may, therefore, be considered a measure of the skewness of a distribution except, of course, m_1 which is necessarily zero for any distribution-symmetrical or otherwise. The simplest of these measures is m_3 .

9.4 Measures of Skewness

Skewness is a measure of the asymmetry of a probability distribution. A distribution is symmetric if it looks the same to the left and right of its center (typically the mean). Skewness quantifies how much a distribution differs from a normal distribution in terms of symmetry.

9.4.1 Pearson's Coefficient

An alternative measure of Skewness is obtained from the relative positions of the mean and the mode in a distribution.

In a symmetrical distribution, the mean, median and mode (assuming the distribution to be uni-modal) coincide. If the distribution is: skewed positively, then

$$\text{mean} > \text{median} > \text{mode}.$$

and if it is negatively skewed, then

$$\text{mean} < \text{median} < \text{mode}.$$

Hence the difference (mean mode) divided by the s.d., is taken as a measure of skewness.

$$Sk = \frac{\bar{x} - M_0}{s} \dots \dots \dots (2.3)$$

This is known as Pearson's first measure of skewness, provided $s > 0$.

Since it is difficult to estimate the mode from a frequency distribution, the empirical relation is used to get another measure of skewness viz.

The moments are not unit free quantity to make it unit free Karl Pearson defined the following four coefficients based on first four central moments.

$$\beta_1 = \frac{m_3^2}{m_2^3}, \quad \beta_2 = \frac{m_4}{m_2^2}, \quad \gamma_1 = \sqrt{\beta_1} \text{ and } \gamma_2 = \beta_2 - 3 \dots \dots \dots (2.1)$$

To get the unit free measurement of skewness is thus defined as

$$g_1 = \frac{m_3}{m_2^{3/2}} \dots \dots \dots (2.2)$$

$\gamma_1 = |g_1|$ is absolute measure of skewness.

An alternative measure of Skewness is obtained from the relative positions of the mean and the mode in a distribution.

In a symmetrical distribution, the mean, median and mode (assuming the distribution to be uni-modal) coincide. If the distribution is: skewed positively, then

$$\text{mean} > \text{median} > \text{mode}.$$

and if it is negatively skewed, then

$$\text{mean} < \text{median} < \text{mode}.$$

Hence the difference (mean mode) divided by the s.d., is taken as a measure of skewness.

$$Sk = \frac{\bar{x} - M_0}{s} \dots \dots \dots (2.3)$$

This is known as Pearson's first measure of skewness, provided $s > 0$.

Since it is different to estimate the mode from a frequency distribution, the empirical relation is used to get another measure of skewness viz.

$$Sk = \frac{3(\bar{x} - M_i)}{s} \dots \dots \dots (2.4)$$

Which is known as Pearson's second measure of skewness. If mean = median = mode, then $S_k=0$.

Check your progress

1. Explain the Skewness.
2. What do you mean by Pearson's coefficient?

9.4.2 Bowley's Coefficient

The measure (2.4) can vary between -3 and 3. The same may be said to be approximately the case with (2.3) because of the empirical relation which is valid for moderately skew distribution. A fourth measure of skewness is obtained by considering the relative positions of the three quartiles of a frequency distribution. For a symmetrical distribution the lower and upper quartiles are equidistant from the median; for a positively skew distribution the lower quartile is nearer the median than the upper quartile is, while for a negative skew distribution the upper quartile is nearer.

Thus $(Q_3 - Mi) - (Mi - Q_1)$ may be taken as a measure of skewness. Its expressed as a pure number on being divided by $(Q_3 - Mi) + (Mi - Q_1) = Q_3 - Q_1$

Which is assumed to be non-zero.

Thus the new measure is

$$Sk = \frac{(Q_3 - Mi) - (Mi - Q_1)}{Q_3 - Q_1} \dots \dots (2.5)$$

This is known as Bowley's measure of skewness. As regards (2.5), it has the limit -1 and 1.

9.4.3 β and γ

The moments are not unit free quantity to make it unit free Karl Pearson defined the following four coefficients based on first four central moments.

$$\beta_1 = \frac{m_3^2}{m_2^3}, \quad \beta_2 = \frac{m_4}{m_2^2}, \quad \gamma_1 = \sqrt{\beta_1} \text{ and } \gamma_2 = \beta_2 - 3 \dots \dots \dots (2.1)$$

To get the unit free measurement of skewness is thus defined as

$$g_1 = \frac{m_3}{m_2^{3/2}} \dots \dots \dots (2.2)$$

$\gamma_1 = |g_1|$ is absolute measure of skewness. The measure given by equation (2.2) can theoretically assume any value between $-\infty$ and ∞ but in practice its numerical value is rarely very high.

For symmetrical distribution $\beta_1 = 0$. $\beta_1 > 0 \rightarrow$ distribution is +1 rely skewed. $\beta_2 < 0 \rightarrow$ distribution is -1 rely skewed.

9.4.4 Another measure based on moments

May be obtained from the Pearson's system of curve. It is defined as –

$$S_k = \frac{\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)} \dots \dots \dots (2.6)$$

It is without sign, It has a draw that it has no limits. If $S_k = 0$ then $\beta_1 = 0$ or $\beta_2 = -3$.

$$\beta_2 \neq -3.$$

$$\text{Since } \beta_2 = \frac{\mu_4}{\mu_2^2}.$$

$$\text{Hence } S_k = 0, \text{ if } \beta_1 = 0.$$

Thus, for a symmetrical μ_2^2 distribution,

$$\beta_1 = 0.$$

9.5 Kurtosis

Another method of describing a frequency distribution is to specify its degree of peakedness or kurtosis. Two distributions 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.

9.5.1 Measure of Kurtosis β_2 and γ_2

This feature of the frequency distribution is measured by

$$\beta_2 = \frac{m_4}{m_2^2}, \quad \text{and} \quad \gamma_2 = \beta_2 - 3 \dots \dots \dots (2.7)$$

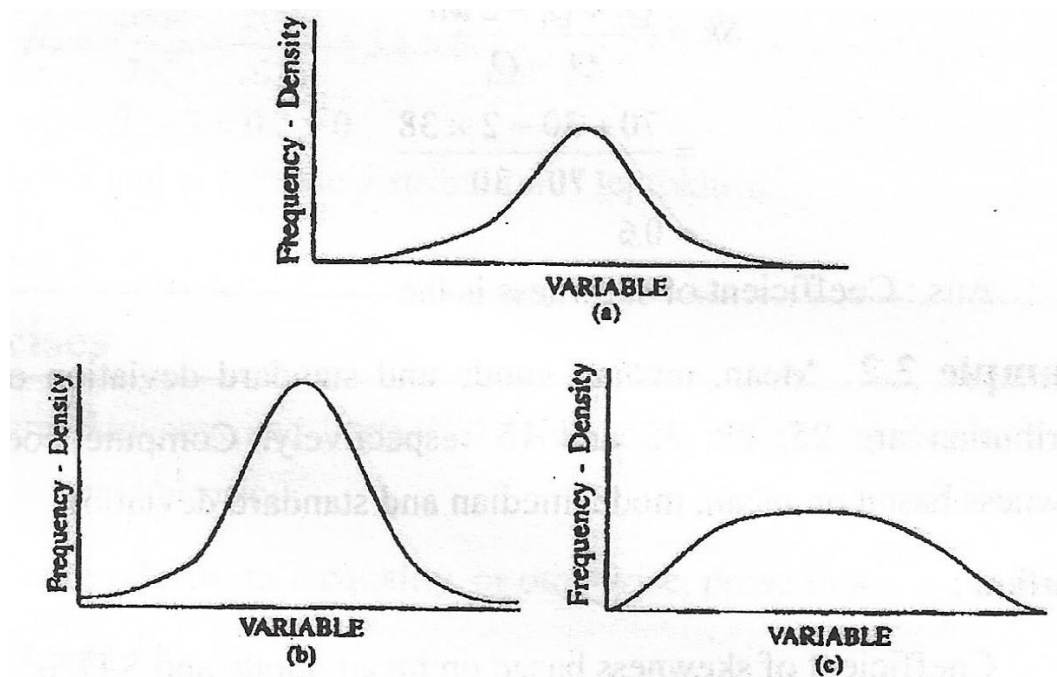


Fig. Three symmetrical distributions with different degrees of kurtosis : (a) mesokurtic, (b) leptokurtic, (c) platykurtic.

Obviously, it is a pure number. For a normal distribution, $\beta_2 = 3$ and $\gamma_2 = 0$.

A positive value of indicates that the distribution has high concentration of value near the central tendency and has high tails, in comparison with a normal distribution with the same standard deviation.

In the same way, a negative value γ_2 means that the distribution has low.

$\beta_2 = 3$ implies that $\gamma_2 = 0$, the kurtosis is same as that of normal curve. The curve is mesokurtic.

$\beta_2 > 3 \rightarrow \gamma_2 > 0$ the Kurtosis is said to be positive and curve called the leptokurtic.

$\beta_2 < 3 \rightarrow \gamma_2 < 0$ the Kurtosis is said to be negative and curve called the platykurtic.

Concentration of values in the neighborhood of the central tendency and how tails, compared to a normal distribution with the same standard deviation. A normal curve is said to be mesokurtic (i.e. having medium kurtosis). A distribution with positive γ_2 is called leptokurtic, and one with negative γ_2 is known as platykurtic. The quantities β_1 and β_2 themselves are sometimes used as measures of skewness and kurtosis, respectively.

That the fourth central moment (m_4) may be used in measuring kurtosis becomes obvious from the fact that higher the kurtosis, the higher will be the effect of the large deviations (from the mean) in the tails when raised to the fourth power. Division of m_4 by s^4 makes the measure a pure number.

Actually, however, β_1 (or γ_2) will be appropriate as a measure of kurtosis or peaked-ness 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.

9.6 Some Solved Examples

Example.1 In a frequency distribution $Q_1=30$, $Q_3=70$ and median is 38, compute coefficient of skewness.

Solution: Coefficient of skewness based on quantities is

$$Sk = \frac{Q_3 + Q_1 - 2Mi}{Q_3 - Q_1}$$

$$= \frac{70 + 30 - 2 \times 38}{70 - 30} = 0.6$$

Answer: Coefficient of skewness is 0.6

Example.2 Mean, median and mode and standard deviation of frequency distribution are 25, 27, 35 and 15 respectively. Compute coefficients of skewness based on mean, mode, median and standard deviations.

Solution: Coefficient of skewness based on mean, mode and S.D. is

$$Sk = \frac{\bar{x} - M_0}{s} = \frac{25 - 35}{15} = \frac{-10}{15} = -0.67$$

Coefficient of skewness based on mean, median and S.D. is

$$Sk = \frac{3(\bar{x} - M_i)}{s} = \frac{3(25 - 27)}{15} = \frac{-6}{15} = 0.4$$

Example.3 First three central moments of a variable are 0, 16 and -64 respectively. Compute coefficient of skewness based on moments and comment.

Solution: Coefficient of skewness based on moments are:

$$\beta_1 = \frac{m_3^2}{m_2^3} = \frac{(-64)^2}{16^3} = 1$$

$$\gamma_1 = \sqrt{\beta_1} = 1$$

Since m_3 is negative, the distribution is negatively skewed coefficient of skewness with sign is

$$g_1 = \frac{m_3}{s^3} = -1$$

Example.4 The standard deviation of a symmetrical distribution is 5 and fourth central moment is 2000. Compute β_1 and γ_2 and comment on the kurtosis of the distribution.

Solution: Since distribution is symmetric

$$m_3 = 0, m_2 = (S.D)^2 = 25, m_4 = 2000$$

then
$$\beta_1 = \frac{2000}{25^2} = \frac{2000}{625} = 3.2 > 3$$

$$\gamma_2 = \beta_2 - 3 = 0.2 > 0$$

Since $\beta_2 > 3$ and $\gamma_2 > 0$; the distribution is leptokurtic.

9.7 Exercises

2.1 What are skewness and kurtosis? Give some suitable measures for skewness and kurtosis.

2.2 Using Cauchy-Schwarz inequality, or otherwise, prove that

(i) $\beta_2 > 1$ and (ii) $\beta_2 - \beta_{21} - 1 \geq 0$.

2.3 Show that the measure of skewness given by (2.4) must lie between -3 and 3 and that the measuring given by (2.5) must be between -1 and 1.

2.4 Prove, by a geometrical argument, that for a J-shaped distribution with its longer tail towards the higher values of the variable, the median is nearer to the first quartile than to the third (A similar argument can be used to show that for the other type of J-shaped distribution, the median is nearer to the third quartile than to the first).

2.5 Consider any symmetrical frequency distribution for a discrete variable. Show that its central moments of odd orders must all be zero.

2.6 The first three moments about 4 of 10 observations were 5.5, 38.5 and 302.5. the 4th moment about 2 of the same 10 observations was 6089.3 It was found later that an observation of 3 was wrongly read as 8. Find the corrected mean, second, third and fourth central moments and measures of skewness and kurtosis.

2.7 In a certain distribution, mean=45 units, median= 48 units, coefficient of skewness = 0.4. The person who supplied the data failed to give the value of the s.d. Estimate it from the above data.

2.8 In a certain distribution the coefficient of skewness based on quarties is 0.6 if the sum of the third first quartile is 100 units and the median is 38 units, find the first and third quartiles.

2.9 Compute $\bar{x}$, s , m_3 and m_4 for the data on length of ear-hear given in Exercise 6.16.

2.10 The scores in English of 250 candidates appearing at an examination have mean= 93.72, $m_2= 97.80$, $m_3 = -114.18$ and $m_4 = 28.396, 14$,

It is later found on security that the score 61 of a candidate has been wrongly recorded as 51. Make necessary corrections in the given values of the mean and the central moments.

2.11 Particulars relating to the monthly wage distribution of two manufacturing firms are given below:

	Firm A	Firms
Mean wage	Rs. 1,477	Rs. 1,495
Median wage	Rs. 1,389	Rs. 1,354
Modal wage	Rs. 1,350	Rs. 1,312
Quartiles	Rs. 1,278 and 1,422	Rs. 1,262 and 1,435
Standard deviation	Rs. 87	Rs. 99

Compare the two distributions.

2.12 The S.D. of a symmetrical distribution is 4. What must be the value of fourth moment about mean so that the distribution be (a) leptokurtic (b) mesokurtic and (c) platykurtic.

9.8 Answer/Suggestions

P-2.6 Partial answers: Corrected first and second moments are 0 and 346.

P-2.7 22.5 unit

P-2.8 30 units

P- 2.9 $\bar{x} = 9.9$, $s = 0.91$, $m_3 = -0.061$ and $m_4 = 29165.60$ (in proper units)

P-2.10 mean= 39.76, $m_2= 99.10$, $m_3= -9.27$ and $m_4 = 39165.60$

P- 2.11

P- 2.12 (a) $\mu_4 < 756$ (b) $\mu_4 < 756$ (c) $\mu_4 > 756$

9.9 Summary

In this unit the properties of symmetrical and asymmetrical distributions have been studied. The measures of skewness have been defined and their interpretations have been given. The peakedness of the frequency curve is explained. The measures of peakedness have been obtained in terms of Central moments.

9.10 Further Readings

1. Kenney, J.F and Keeping, E.S.: Mathematics of Statistics, Part I (Ch. 7) Van Nostrand, 1954 and Affiliated East- West Press.
2. Mills, F.C.: Statistical Methods (Ch.5) H. Holt, 1955.
3. Yule G.U. and Kerdall, M.G.: Introduction to the Theory of Statistics (Ch.6) Charles Griffin, 1953.
4. Fundamentals of Statistics volume I by Goon, Gupta and Dasgupta.



**U. P. Rajarshi Tandon
Open University**

Master of Science PGMM -108N Mathematical Statistics

Block

4 Correlation and Regression

Unit- 10 Bivariate Data and Correlation	197-236
Unit- 11 Regression	237-252
Unit- 12 Line of Regression	253-282
Unit- 13 Correlation and Intra Class Correlation	283-306
Unit- 14 Theory of Attributes	307-332

Block-4

Correlation and Regression

The ***Block - 4 Correlation and Regression (Two Variables and Association)*** deals with correlation and Regression (two variable cases only) along with theory of Attributes. It consists of four units. The *first unit* of this block describes the concept of correlation, scatter diagram and properties of correlation coefficient. The *second unit* of this block discusses concept of regression and the *third unit* deals with regression lines and their properties. The fourth *unit* of this block defines Spearman's rank correlation coefficient and interclass correlation coefficient. The *last unit* describes the association of attributes, independence, contingency table, measures of association and Yates correction.

At the end of every block/unit the summary, self-assessment questions and further readings are given.

Unit-10 : Bivariate Data and Correlation

Structure

10.1 Introduction

10.2 Objectives

10.3 Scatter Diagram

10.4 Karl Pearson's Coefficient of Correlation

10.5 Properties of Correlation Coefficient.

10.6 Limits of Correlation Coefficient

10.7 Effects of change of origin and scale on the correlation coefficient

10.8 Exercises

10.9 Answers

10.10 Summary

10.11 Further Readings

10.1 Introduction

Bivariate Distribution: Distribution involving two discrete variables is called a Bivariate distribution. For example,

1. The height and weights of the students of a class in school.
2. The daily petrol used by a scooter owner and the mileage covered by it.

Frequency. Let (x_i, y_i) , $i = 1, 2, 3, \dots, m$; $j = 1, 2, \dots, n$, be a bivariate distribution. If the pair (x_i, y_i) occurs f_{ij} times then f_{ij} is called the frequency of the pair (x_i, y_i) and $\sum_{i=1}^m \sum_{j=1}^n f_{ij} = N$ = the total frequency.

This unit concerns with the relationship between two variables. In general, two variables are said to be related if they vary together in a systematic fashion. If two variables change in the same direction (either both increase or both decrease) they are said to be positively related. If the variables change in opposite direction (as one increases, the other decrease and vice versa), they are said to be negatively correlated. For example, variables level of education and level of income are expected to be positively related while level of education and level of prejudice are negatively related also. The weight of an adult depends to some extent on the height. Similarly, there are various functional forms or shapes in which two variables might be related. However, in this unit we will be concerned with linear relationship only between the variables— those relationships that can be represented by a straight line. For example, a linear relationship between the dependent variable Y and the independent variable X is of the form: $Y = a + bx$ where ‘ a ’ is the intercept and b is the slope of the line. While studying a bivariate relationship, several questions are interest: Are the two variables related? Is the relationship linear or nonlinear? How strong is the relationship? etc., What is the nature of the relationship? Various statistical techniques such as scatter plot, correlation, and regression analysis are used to answer these questions.

10.2 Objectives

After going through this unit you should be able to-

- Discuss bivariate data and scatter diagram.
- Define correlation Coefficient.
- Describe the properties of correlation coefficient.

10.3 The Scatter Diagram

The simplest mode of diagrammatic representation of bivariate data is the use of *scatter diagram* (or dot diagram). Taking two perpendicular axes of co-ordinates, one for x and the other for y, each pair of values $\{(x_i, y_j), i=1,2,3,\dots,n\}$ is plotted as plotted as a point on graph paper or xy-plane. The whole set of each (x_i, y_j) is represented as a point taken together constitutes the scatter diagram or ‘dot’ diagram. After obtaining the raw data on two quantitative variables, these data must first be arranged and paired as $(x_i, y_j, i=1,2,\dots,N)$. It is not efficient to make a contingency table because both variables can potentially take a large number of values. Instead these data are arranged in the form of a graph known as Scatter Diagram. A scatter diagram graphically summarizes the data on two quantitative variables by showing the joint distribution of the values on two variables. Each point in a scatter plot represents individual’s scores on the two variables represented by the two axes of the graph. A given case’s point is located at the intersection of that case’s values on each variable. To construct a scatterplot: (i) Draw an X-axis or horizontal axis and label it with the name and values of the independent variable. (ii) Draw a Y-axis or vertical axis and label it with the name and values of the dependent variable. To keep the graph from appearing either too flat or too steep, keep the height of the vertical axis equal to two thirds the length of horizontal axis. (iii) Plot the pairs of points (x_i, y_j) as dots between the two axes. For larger data sets, use SPSS to construct scatter plots.

Example 1.1: A researcher is interested in studying the relationship between level of education (measured in years) and income (measured in thousands of rupees). Let level of education be the independent variable denoted by X and income be the dependent variable denoted by Y. For a sample of 15 people, the data are given below:

Case # :	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X:	7	12	8	12	14	9	18	14	8	12	17	10	16	10	13
Y:	9	16	14	12	11	16	19	13	13	14	14	16	15	10	18

For these data, make a scattered diagram.

Solution: A scattered diagram of data on income and education is shown by Fig. 1.1 on next page.

10.3.1 Reading and Interpreting a Scatter Diagram.

A scatter diagram provides several important pieces of information about the relationship between X and Y. In fact, scatter diagram can often provide information about the relationship that is missed from a casual inspection of the statistical (numerical data) usually computed in connection with a correlation analysis. For this reason, a scattered diagram should always be constructed as the first step in correlation analysis. Using a scatter diagram one can address the following questions:

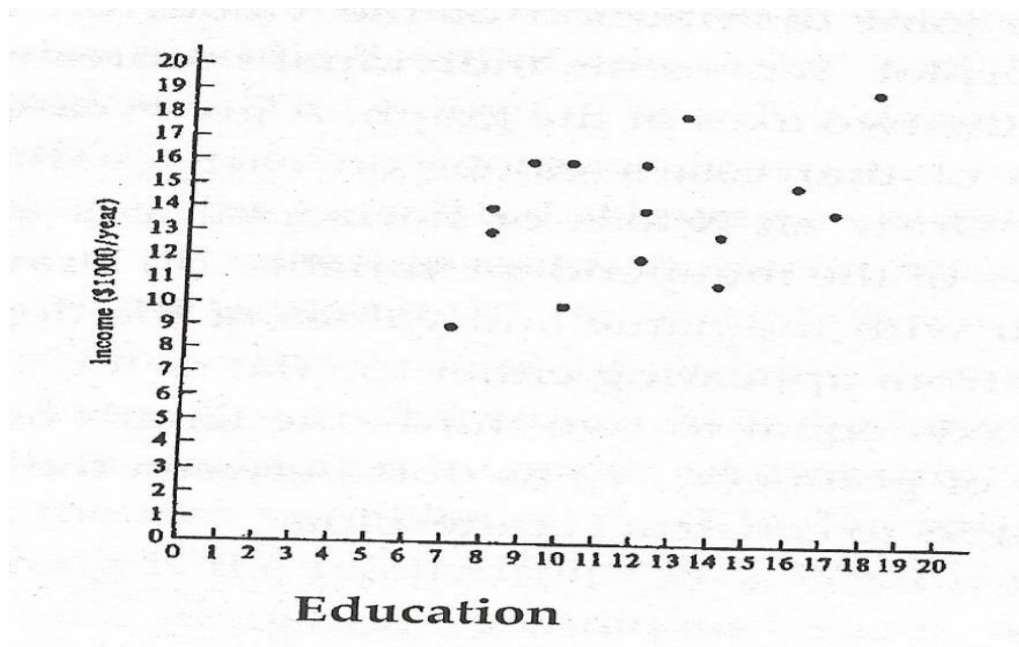
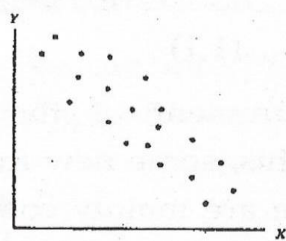


Figure 1.1 An Escambled Scatter Diagram

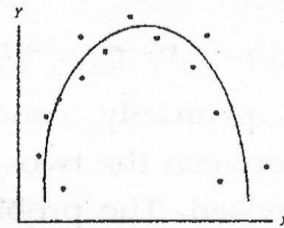
- (a) Is the relationship linear? If an examination of the scatter diagram suggests a football of egg-shaped pattern, two variables are likely to be related in a linear manner.
- (b) Does there exist a relationship? If the football is tilted, two variables are expected to be related. If the football is either standing vertically or hanging horizontally, if the points are parallel to either of the axes, then there are no relationships between the two variables since the change in one variable will not imply a change in the other variable.

- (c) How strong is the relationship? If all the points fall on a straight line (that is, there is no scatter of points), the relationship is perfect. And if the pattern in the scatter is either circular or random, or nonlinear, then there is no linear relationship between X and Y. The thinner the relationship. Scatter diagram reveals the extent to which variables covary, and the amount of variability found in each variable considered singly.
- (d) What is the nature of the relationship? If the football is tilted from left to right, it indicates a positive relationship and if the football is tilted from right to left it is indicative of a negative relationship.



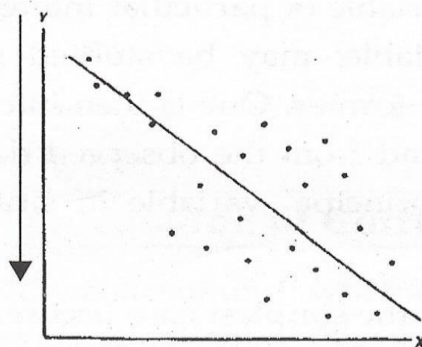
Linear, Negative, Moderate

Fig. 1.2a



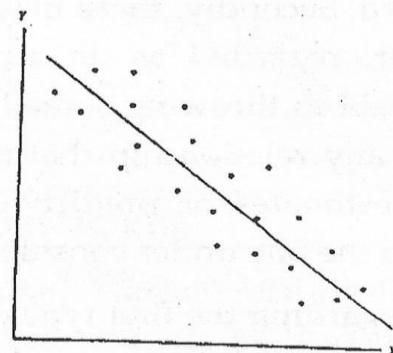
Nonlinear, inverted-U shape

Fig. 1.2 b



Heteroscedastic

Fig. 1.2c



Homoscedastic

Fig.. 1.2d

- (e) Are there any outliers in the data? An outlier is a case that is radically different from the majority of other cases in terms of its combined or joint values on X and Y. Although an outlier may show scores on X and Y that are each well within the normal ranges for those variables the pair of scores may make the case quite deviant. In a scatter diagram the outliers will appear conspicuously removed from the rest of the points. Identifying outliers is important for several reasons. First, an outlier may represent a case whose scores on X and or Y have been recorded incorrectly. Second outliers may show deviant combined

scores on X and Y because they did not understand (or deliberately disobeyed) instructions given during data collection. Third, outliers represent cases to which statements about the relationship between X and Y drawn from the majority of cases do not apply. Fourth, outliers exert a disproportionate effect on computed correlations.

10.4 Karl Pearson's Correlation Coefficient

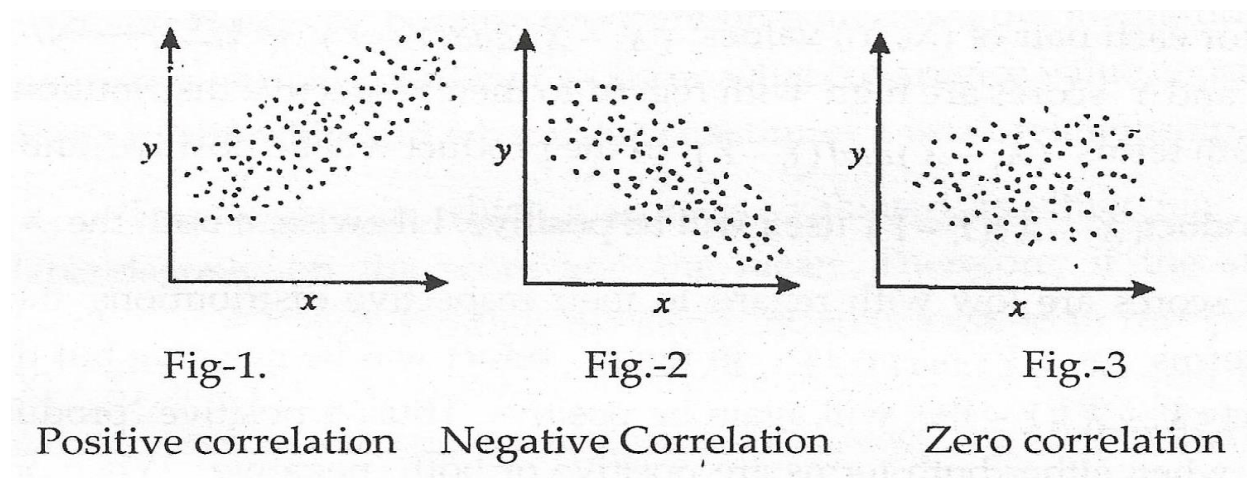
An examination of the scatter diagram in example 1.1 suggests a linear, positive and moderately strong relationship between education and income. But examination of a scatter diagram is only visual, approximate and subjective way of measuring the relationship even the change of scales of axes has an effect on the slope of the curve. We would like to measure the relationship in a more accurate, concrete and objective manner. That is we would like to quantify the scatter diagram. A measure used to qualify the degree of linear relationship between two variables is called Pearson product moment correlation coefficient, denoted by r . Just as observed in graphed frequency distributions, Pearson r gives a more precise indication of the linear relationship between X and Y than is available from inspection of a scatter diagram. Pearson r is a symmetric statistic that is its value and nature does not depend upon whether X is independent and Y is dependent or vice versa. The correlation between X and Y is equal to correlation between Y and X. That is

$$r_{xy} = r_{yx} = r \dots \dots \dots (1.1)$$

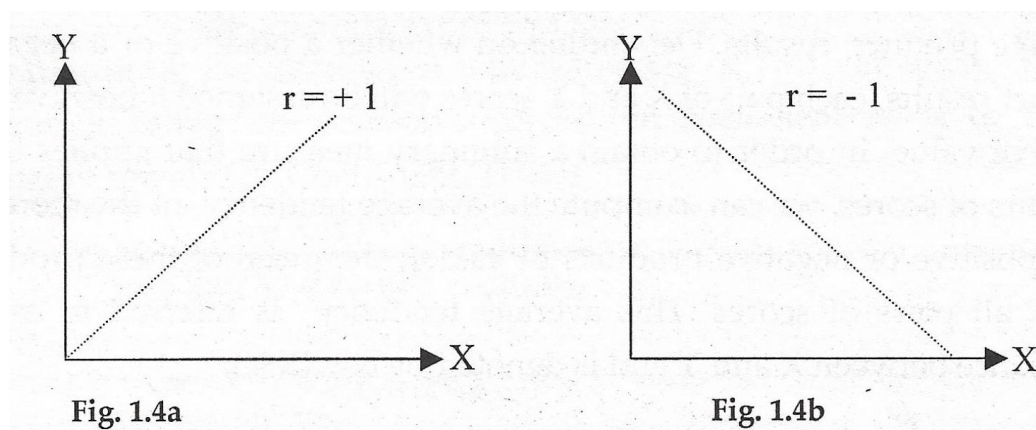
here we are primarily concerned with measurement of the linear relationship between the two variables and for this some new methods have to be devised. The problems with which we are mainly concerned may be of two types. First, the data may reveal some relationship between the two variables and we may want to measure the extent to which they are related. Secondly, there may be one variable, may be studied for its possible aid in throwing some light on the former. One is then interested in using any relationship that may be found from the observed data for making estimates or predictions of the principal variable in situations similar to the one under consideration.

Regarding the first type of problem the simplest case occurs when from the scatter diagram or otherwise, the variables are found to be linearly related, at least approximately. If it is found that as

one variable increases the other also increases, in general or on the average there will be said to be *positive correlation* between them. This will be the case, for example, when the data relate to the height and weight of people or the score in mathematics and the score in statistics of students in a college. On the other hand, as one variable increases, the other may decrease on the average. We then say that there is *negative correlation* between them. There may be a third situation where as one variable increases, the other remains constant on the average. This is the case of zero or no correlation and the two variables are then said to be uncorrelated. A near zero correlation is expected when we have data on height and IQ of students in HS institution. Following scatter diagrams show the nature of correlation between two variables x and y .



We can interpret the way scatter of cluster as the properties of relationship between the two variables.



10.4.1 Calculation of Correlation Coefficient

In determining whether a systematic linear relationship exists between two variables, one seeks to find out whether high scores on one variable are paired with high scores on the other variable and low scores are paired with low scores or whether high scores on one variable are paired with low scores on the other variable and low scores are paired with high scores or neither. In order to label a particular X scores as either high or low relative to the other X scores, we compare each particular Y scores to other Y's with the mean $\bar{Y}$. Of all Y scores. If the X score is high, then it will be higher than $\bar{X}$ and the difference $X - \bar{X}$ will be positive. If the X score is low, than X and the difference $X - \bar{X}$ will be negative. Likewise, for the Y scores. In this way we calculate all the deviations $X_i - \bar{X}$ and $Y_i - \bar{Y}$.

To determine whether a high score on X is paired with a high score on Y, and vice versa form the product $(X_i - \bar{X})(Y_i - \bar{Y})$ of the two deviations terms for each pair of (X_i, Y_i) values $(X_i - \bar{X})$ and $(Y_i - \bar{Y})$ $i=1,2,3,\dots,N$. If both X and Y scores are high with regard to their respective distributions. Then both terms $(X_i - \bar{X})$ and $(Y_i - \bar{Y})$ in the product will be positive and the product $(X_i - \bar{X})(Y_i - \bar{Y})$ itself will be positive. Like wise, if both the X an Y scores are low with regard to their respective distributions, then both terms $(X_i - \bar{X})$ and $(Y_i - \bar{Y})$ in the product will be negative but the pr $(X_i - \bar{X})(Y_i - \bar{Y})$ oduct will again be positive. Thus a positive product results when either both terms are positive or both negative. When one (not both) of the differences is negative, signifying either a high X score paired with a low Y score or a low X score paired with high Y score negative product results. Depending on whether a positive or a negative product results each pair of X and Y scores will be assigned a positive or a negative value. In order to obtain a summary measure that applies to all the pairs of scores, we can compute the average tendency of the scores to have positive or negative products by taking the mean of these products across all pairs of scores. This average tendency is referred to as the covariance between X and Y is denoted by $cov(X,Y)$; $\mu_{12} = cov(X,Y)$

$$= \frac{(X_1 - \bar{X})(Y_1 - \bar{Y}) + (X_2 - \bar{X})(Y_2 - \bar{Y}) + \dots + (X_n - \bar{X})(Y_n - \bar{Y})}{N}$$

$$= \frac{1}{N} \sum_{i=1}^N [(X_i - \bar{X})(Y_i - \bar{Y})]$$

$$= \frac{1}{N} \sum x_i y_i \dots \dots \dots (1.2)$$

Where $x_i = X_i - \bar{X}$ and $y_i = Y_i - \bar{Y}$ for $1, 2, 3, \dots, N$

The covariance measures how X and Y vary together. If the covariance between X and Y is positive, a positive linear relationship exists between X and Y. $\text{Cov}(X, Y) > 0$ implies that positive $x_i y_i$'s dominates negative $x_i y_i$'s. If the covariance between X and Y is negative, a negative linear relationship exists between X and Y. If the covariance between X and y is 0, no linear relationship exists between X and Y and the sum of positive $x_i y_i$'s equals in magnitude to sum of negative $x_i y_i$'s help us to interpret the strength of the relationship between the two variables. However, because there are not bounds on the magnitude of the covariance term, it is difficult to know what covariance value constitutes a strong relationship and what value constitutes a weak relationship.

The manner in which a score is evaluated as either high or low depends only on the score and the mean. Therefore, if the standard deviations of X and Y differ, the same difference value in raw points on $(X_i - \bar{X})$ and $(Y_i - \bar{Y})$ would not in general mean the same real distance above or below the respective means. And since we are comparing two distributions of scores pairwise, it is important to standardize the differences so that it means the same things in both distributions. This standardizing can be done in various ways. One way is to divide the total covariance by the product of total variance of two variables. What we obtain is called the Pearson's correlation coefficient r_{xy} is or product moment correlation coefficient. That is,

$$r_{xy} = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}} = \frac{\mu_{12}}{\mu_{11}\mu_{22}} \dots \dots \dots (1.3)$$

$$r_{xy} = \frac{\frac{1}{N} \sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\left[\frac{1}{N} \sum (X_i - \bar{X})^2\right] \left[\frac{1}{N} \sum (Y_i - \bar{Y})^2\right]}}$$

$$r_{xy} = \frac{\sum [(X_i - \bar{X})(Y_i - \bar{Y})]}{\sqrt{\sum (X_i - \bar{X})^2} \sqrt{\sum (Y_i - \bar{Y})^2}} \dots \dots \dots (1.4)$$

Again $N \text{ cov}(X, Y) = \sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum X_i Y_i - N \bar{X} \bar{Y}$

$$= \sum X_i Y_i - \left(\sum X_i \right) \left(\sum Y_i \right) / N \dots \dots \dots (1.5)$$

$$\begin{aligned} N \text{ var } (X) &= \sum (X_i - \bar{X})^2 = \sum X_i^2 - N \bar{X}^2 \\ &= \sum X_i^2 - \frac{(\sum_i X_i)^2}{N} \end{aligned} \quad (1.6)$$

$$\text{And similarly, } N \text{ var } (Y) = \sum Y_i^2 - (\sum Y_i)^2 / N \dots \dots \dots (1.7)$$

Hence r may be expressed in the alternative forms;

$$r_{xy} = \frac{\sum [x_i y_i / N] - \bar{X} \bar{Y}}{\{[\sum_i X_i^2 / N] - \bar{X}^2\}^{1/2} \{[\sum_i Y_i^2 / N] - \bar{Y}^2\}^{1/2}} \dots \dots \dots (1.8a)$$

$$= \frac{N \sum_i X_i Y_i - (\sum_i X_i)(\sum_i Y_i)}{\{N \sum_i X_i^2 - (\sum_i X_i)^2\}^{1/2} \{N \sum_i Y_i^2 - (\sum_i Y_i)^2\}^{1/2}} \dots \dots \dots (1.8)$$

The form (1.4) will be found to be the most convenient for remembering the def; Either Form (1.8) or more convenient for computation work. If any one of $\sum X_i Y_i$, $\sum X_i^2$ or $\sum_i Y_i^2$ is quite large in magnitude, then we use (1.8a) computing from raw data.

These formula require the computation of only five different sums:

$$\sum X_i, \sum Y_i, \sum X_i Y_i, \sum X_i^2, \sum Y_i^2$$

Example 1.2: Using the data in Example 1.1 calculate r.

Solution: The necessary information is presented in the following table. The first two columns contain the given X and Y score.

Table 1.1

X_i	Y_i	$X_i Y_i$	X_i^2	Y_i^2
7	9	63	49	81
12	16	192	144	256
8	14	112	64	256

12	12	144	144	144
14	11	154	196	121
9	16	144	81	256
18	19	342	324	361
14	13	182	196	169
8	13	104	64	169
12	14	168	144	196
17	14	238	289	196
10	16	160	100	256
16	15	240	256	225
10	10	100	100	100
13	18	234	169	324
Total	180	210	2577	3050

Putting $N=15$, $\sum X=180$, $\sum Y=210$, $\sum X^2=2320$, $\sum Y^2= 3050$, $\sum XY=2577$ in equation (1.8) and (1.1), we get

$$r_{xy} = \frac{N(\sum X_i Y_i) - (\sum X_i)(\sum Y_i)}{\sqrt{[N(\sum X_i^2) - (\sum X_i)^2] \{N(\sum Y_i^2) - (\sum Y_i)^2\}}}$$

$$r = r_{xy} = r_{yx} = \frac{15(2577) - (180)(210)}{\sqrt{[(15 \times 2577 - 180^2)(15 \times 2320 - 210^2)]}} = 0.43$$

As a measure of degree of relationship between linearly related variable X & Y.

Check your progress

1. Explain the Scatter Diagram.
2. What do you mean by Correlation Coefficient?

10.4.2 Interpreting (r)

Sign and magnitude of r: The Pearson correlation tells two things about a relationship between X and Y. First, the sign or r (Positive or Negative) indicates the direction of the relationship. Secondly the absolute magnitude of r reflects the strength of the linear relationship between X and Y. Correlations can range in value from -1 to 0 to +1. An $r=.43$, indicates a linear positive moderate relationship between education and income.

10.4.3 Calculation of Correlation Coefficient from Grouped Data

Suppose the values of X and Y are given in the form of a bivariate frequency table with m classes for X and N classes for Y. Let us denote by X_i the mid-point of the i th class of X and X_i = mid-point of i th class, Y_j =mid-point of j th class, f_{ij} = frequency of the (I,j)the cell., $\sum_i \sum_j f_{ij} = N = \sum_i f_{i0} = \sum_i f_{ij}$, $i = 1,2,3, \dots m$, $j = 1,2,3, \dots n$.

Where $f_{i0} = \sum_j f_{ij}$ $f_{0j} = \sum_i f_{ij}$

and $N = \sum_{i,j} f_{ij}$ N is used for no. of classes for y.

Obviously,

$$\bar{X} = \sum_i \frac{X_i f_{i0}}{N} = \sum_i \frac{X_i f_{0i}}{N} \dots \dots \dots (1.11)$$

To reduce computational labour, we may take changes of base and scale for either x or y both x and y. It will be found advantageous (when the classes for each variable are equally wide) to take as bases two class marks, say A and B, somewhere in the middle of the ranges of x and y, respectively, and as units the widths of the corresponding class intervals, say c and d.

The new variables are then $u = (X-A)/c$ and $v = (Y-B)/d$. It follows that

$$r = \frac{\sum_{i,j} (u_i - \bar{u})(v_j - \bar{v}) f_{ij}}{\{\sum_i (u_i - \bar{u})^2 f_{i0}\}^{1/2} \{\sum_j (v_j - \bar{v})^2 f_{0j}\}^{1/2}} \dots \dots (1.12)$$

$$= \frac{\sum_i u_i v_i f_{ij} - n \bar{u} \bar{v}}{\{\sum_i u_i^2 f_{io} - n \bar{u}^2\}^{\frac{1}{2}} \{\sum_i v_i^2 f_{io} - n \bar{v}^2\}^{\frac{1}{2}}} \dots \dots \dots (1.13)$$

$$\text{Since } \bar{u} = \sum_i \frac{u_i f_{io}}{N} \text{ and } \bar{v} = \sum_i \frac{v_i f_{io}}{N}$$

It is easy to calculate $\sum_i u_i f_{io}$, $\sum_i u_i^2 f_{io}$, $\sum_j v_j f_{io}$, $\sum_j v_j^2 f_{io}$, from the bivariate frequency table.

Cross classification data are tabulated in the form given in table 1.2. Data are presented in the cross classified form when the number of paired observation on any two variable X and Y is larger. Such a system of tabulation, being more scientific, makes the data more manageable.

Example 1.3: Cross classification of Data on power consumption and Person Employed in 75 Companies Calculator is given in table 1.2.

Table -1.2
Persons Employed (X)

	10-14	15-19	20-24	25-29	30-34	35-39	Total
050-050	5	7					12
060-069	10	4	5				19
070-079		2	6	3			11
080-089			5	8	5		18
090-099				3	6	2	11
100-109					3	1	4
Total	15	13	16	14	14	3	75

meaningful. It is result of classification of data on each of the two variables in an appropriate number of class intervals displays in columns and rows, frequencies entered in the body of the table as cell frequencies.

A cross classification table makes interesting reading. Take the first column and first row. Out of 15 companies which employ between 10 to 14 persons, five companies consume power between 50 to 59 units per hour and ten between 60 to 69 units per hour. Similarly, out of twelve companies

consuming power between 50 to 59 units per hour, five companies employ between 10 to 14 persons and seven between 15 of 19 persons. Likewise for all other rows and columns.

Given the cross classified paired data on the two variables, the coefficient of correlation r between power consumption and the number of persons employed can be found by using.

$$r = \frac{N \sum f d_x^2 - (\sum f d_x)(\sum f d_y)}{\sqrt{[N \sum f d_x^2 - (\sum f d_x)^2] [N \sum f d_y^2 - (\sum f d_y)^2]}} \dots \dots \dots (1.14)$$

Computational Procedure:

The use of the above equation consists of the following steps in the order listed:

- Find the mid point, x and y for each class both for X and Y data series.
- Decide the assumed means A and B for X and Y series, respectively.
- Obtain d_x and d_y in the define above.
- Multiply each d_x and d_y by the corresponding column/row frequency f to get $f d_x$ and $f d_y$ and find the sums
- Take the square of each d_x and d_y to get d_x^2 and d_y^2 and multiply each d_x^2 and d_y^2 by the corresponding class frequency f to get $f d_x^2$ and $f d_y^2$ and the sums
- Obtain the product of each d_x and d_y and multiply by frequency f indicated in the appropriate cells and write them in squares in the left had corner of the concerned cells.

Add these product values over all rows and columns to get and obtain the sum this sum for all columns should be same as the one for all rows.

All values so obtained may be substituted in equation (1.14) to solve for r . these six steps computations are illustrated to get the following sums:

$$\sum f d_x = 8, \sum f d_y = 9, \sum f d_x^2 = 165, \sum f d_y^2 = 170,$$

$$\text{and } \sum \sum f_{ij} = 145, u_i v_i = 145.$$

With $N = \sum f_x = 75$, when the above values are substituted,

we have

$$r = \frac{N \sum f d_x d_y - (\sum f d_x)(\sum f d_y)}{\sqrt{[N \sum f d_x^2 - (\sum f d_x)^2] [N \sum f d_y^2 - (\sum f d_y)^2]}}$$

$$r = \frac{(75)(145) - (9)(8)}{\sqrt{[(75)(170) - (8)^2] [(75)(165) - (9)^2]}} = 0.87$$

10.5 Properties of Correlation Coefficient

The properties of correlation coefficient r are:

- (1) Correlation coefficient is a pure number having no unit.
- (2) The correlation coefficient r lies between -1 to +1, that is $-1 \leq r \leq 1$
- (3) It is not affected by linear transformation of variables, that is, if $u=X-A$, $v=Y-B$
Then $r_{xy}=r_{uv}$ and is independent of A and B .
- (4) If correlation coefficient between X and Y be r and regression coefficient b_{yx} and b_{xy} ,
then $r = \sqrt{b_{yx} b_{xy}}$
- (5) The sign of r , b_{yx} and b_{xy} are same.
- (6) If two variables x and y are independent then $r=0$, but converse is not true.
- (7) r^2 , the square correlation coefficient is referred to as coefficient of determination while $(1-r^2)$ as coefficient of non determination.
- (8) The quantity $\sqrt{1-r^2}$ is referred to as coefficient of alienation. Below some properties are discussed.

Interpretation of values -1, +1 and 0 or r

- (1) If $r=+1$, it means that r.v. Y is proportional to X , which ensure perfect linear relationship between X and Y . In this case all points into scatter diagram lie on a straight line, extending from left bottom to the right bottom, if there is an increase in X then there will be proportional increase in Y .
- (2) If $r=-1$, it means that r.v. Y decreases with the increase in X and vice versa. There is perfect negative relationship between the variables, All the points in the scatter diagram (i.e. on a straight line extending from left top to right bottom).

- (3) The value $r=0$ confirms the lack of linear relationship between two variables. All the points are scattered on the graph and hardly any these points lie in a straight line.

10.6 Limits of the Correlation Coefficient

The correlation coefficient r lies between -1 and $+1$. That is $-1 \leq r \leq +1$ or $|r| \leq 1$. The Pearson correlation coefficient r , unlike the covariance, can only take on values between -1 and $+1$ inclusive. The covariance $\text{Cov}(X, Y)$ may take any value from but the Pearson coefficient r can take on values for -1 to $+1$ both inclusive. A value of -1 indicates a perfect negative linear relationship while a value of $+1$ of r indicates a perfect positive linear relationship between the two variables. In both these cases, all pairs of points fall exactly along a straight line. It should be noted that whenever either s_x or s_y or both are zero, r cannot be calculated because division by 0 is not permitted; s_x and s_y denote standard deviations of the variables X and Y respectively.

In formula (1.3), let us put $x_i = (x_i - \bar{x})/s_x$ and

$$y_i = \frac{(y_i - \bar{y})}{s_y} \dots \dots \dots (1.15)$$

Then $\sum_i x_i^2 = \sum_i y_i^2 = n$

$$\text{and } r = \frac{\sum_i x_i y_i}{n} \quad \text{since}$$

$$\sum_i (x_i + y_i)^2 \geq 0,$$

$$\text{Or } \sum_i x_i'^2 + \sum_i y_i'^2 + 2 \sum_i x_i y_i \geq 0, \dots \dots \dots (1.16)$$

$$\text{Or } 2n(1 + r) \geq 0, \dots \dots \dots (1.17)$$

We have $r \geq -1$.

$$\text{Again } \sum_i (x_i - y_i)^2 \geq 0, \quad (\text{always})$$

$$\sum_i x_i'^2 + \sum_i y_i'^2 - 2 \sum_i x_i y_i \geq 0, \dots \dots \dots (1.18)$$

Or $2n(1 + r) \geq 0$.

Hence $r \leq 1$(1.19)

from (1.17) and (1.18), we have $-1 \leq r \leq 1$ (1.20)

or $|r| \leq 1$(1.21)

This shows that change of Origin and scale on the Correlation Coefficient.

10.7 Effects of change of Origin and scale on the Correlation Coefficient

Let $u = (x - A)/c$ and $v = (y - B)/d$, where A, B, c and d are four arbitrarily chosen constants corresponding to each to pair of values. We have a pair of values of the new variables, i.e.,

$$u_i = \frac{(x_i - A)}{c} \text{ and } v_i = \frac{(y_i - B)}{d} \dots \dots \dots (1.22)$$

Hence, $x_i = A + cu_i$ and $\bar{x} = A + c\bar{u}$,

So that $x_i - \bar{x} = c(u_i - \bar{u})$.

Similarly, $y_i - \bar{y} = d(v_i - \bar{v}) \dots \dots \dots (1.23)$

Hence $cov(X, Y) = cd \times \sum_i \frac{(u_i - \bar{u})(v_i - \bar{v})}{n} = cd cov(u, v)$,

$Var(X) = c^2 \times \sum_i (u_i - \bar{u})^2 / n = c^2 var(u) \dots \dots \dots (1.25)$

Thus, $r_{xy} = \frac{cd cov(u, v)}{\sqrt{c^2 var(u)} \sqrt{d^2 var(v)}} = r_{uv} \dots \dots \dots (1.26)$

Which means that-

- (a) If c and d are of same, then $r_{xy} = r_{uv}$.
- (b) If c and d are opposite sign, then $r_{xy} = -r_{uv}$.

Thus the coefficient of correlation is independent of the change of origin but not of scale.

If two variables X and Y are independent, they are uncorrelated but the converse is not true.

As mentioned above we found that when x and y are independent then will be zero and then $r_{xy}=0$. Thus when X and Y are independent then there is no correlation between X and Y.

However, the converse is not true. It can be proved by an example-

Suppose we have the following data.

x:	1	2	3	4	5	6	7
y:	9	4	1	0	1	4	9

Here the variables x and y are related by the equation $y = (x-4)^2$, but

$$\text{cov}(x, y) = \frac{1}{n} \left[\sum xy - \frac{\sum(x) \sum(y)}{n} \right] = \frac{1}{7} \left[112 - \frac{28 \times 28}{7} \right] = 0$$

Thus $r_{xy} = 0$.

Remark: Even better than r, as a measure of strength, is r^2 , the squared correlation. The r^2 provides a non-linear measure of relationship strength between X and Y. This situation is shown in the figure below.

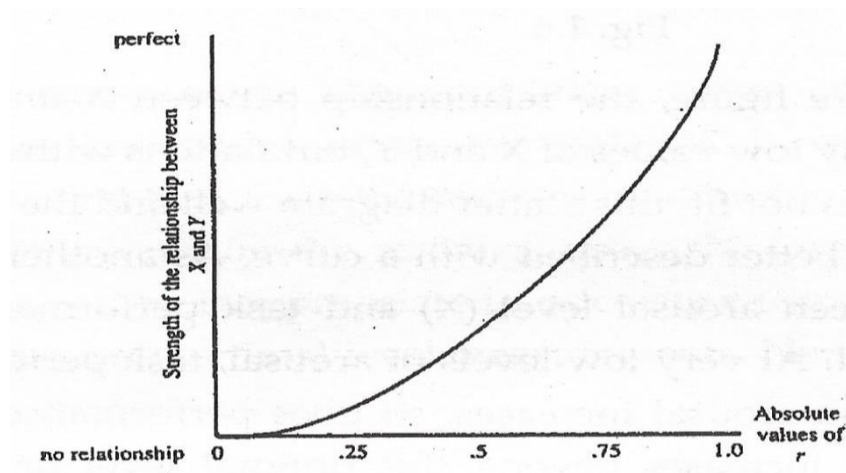


Fig.1.5

The r^2 (the squared correlation, also known as coefficient of determination) provides a linear measure of the strength of the linear relationship between X and Y. This situation is shown in the figure.

Thus, r^2 is a more sensitive value than r . An $r^2=0$ indicates complete absence of any linear relationship between X and Y . In a descriptive sense, an $r=.43$ indicates a moderately strong positive linear relationship between education and income. An $r^2 = 0.43^2 = 0.1849$ indicates, in the predictive sense, that about 18% of the variability in income can be accounted for by education and vice versa.

(ii) Value of r and non linearity: The r is sensitive only to linear relationship between variables. A low correlation may not always indicate that variables are unrelated. It may mean instead that they are not related in a linear fashion, as shown in the following figure.

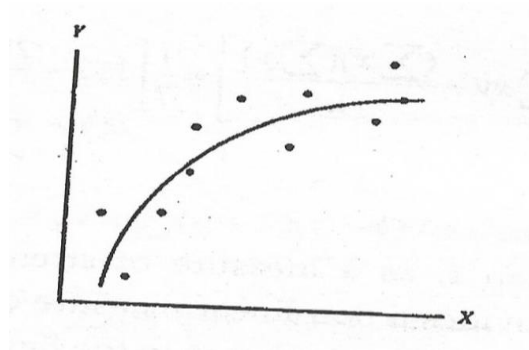


Fig.1.6

In the above figure, the relationship between X and Y increases rapidly at relatively low values of X and Y , but flattens with higher values. A straight line does not fit this scatter diagram well and the value of r will be quite low. It is better described with a curve. As another example, the relationship between arousal level (X) and task performance (Y) is of inverted -U shaped. At very low levels of arousal task performance tends to be very low. As arousal increases so does performance tends point. As arousal increases beyond this optimal level or critical level, performance deteriorates. Deviations from linear relationship are frequent in social sciences and this makes an examination of scatter plot for nonlinearity.

Example 1.4: A professor wishes to determine, for 10 students in his advanced quantitative analysis class, if there is a relationship between the amount of stress a student feels midway through the course and the grade the student receives at the completion of the course. Both variables are assumed to be ordinal level with a higher score on the stress scale indicating greater perceived stress. The data are given below:

Student	Stress (X)	Grade (Y)
1	3	A
2	2	A-
3	5	B
4	3	B+
5	5	B
6	7	B
7	1	A
8	8	C
9	10	D
10	10	C

Solution: Higher stress scores represent more stress, and higher grades represent more knowledge of course material. It will be easier (and appropriate) to first transform the two sets of values to rankings and then compute the Pearson r on the rankings using the formulas discussed in this chapter. It will be inappropriate to calculate r using the original data because they are only ordinal level measurements. Necessary information is presented in the following table.

Student	Original X	Ranked X	Original Y	Ranked Y
1	3	7.5	A	1.5
2	2	9	A-	3
3	5	5.5	B	6
4	3	7.5	B+	4
5	5	5.5	B	6
6	7	4	B	6
7	1	10	A	1.5
8	8	3	C	8.5
9	10	15	D	10
10	10	1.5	C	8.5

Notice that students 9 and 10 have the same original score on stress. So they are tied for first and second place in the rankings. As a result they are assigned the average of these two ranks, 1.5, each. The same procedure is used whenever ties occur.

$$r = \frac{[n \sum X_i Y_i - (\sum X_i)(\sum y_i)]}{\sqrt{[n \sum x_i^2 - (\sum X_i)^2][n \sum y_i^2 - (\sum y_i)^2]}}$$

$$r = \frac{[(10)(226.50) - (55)(55)]}{\sqrt{[10(383.50) - (55)^2][10(382) - (55)^2]}}$$

$$= \frac{2265 - 3025}{\sqrt{(810)(795)}} = -\frac{760}{802.46} = -0.95$$

This result suggests that the greater the perceived stress the lower the grade, and the lower the perceived stress the higher the grade. Notice that the since these ranking contain ties, it will not be appropriate to use Spearman rank correlation coefficient. We should apply some correlation in the formula for?

Example 1.5: In an annual high school English essay contest, eight students submitted essays to be judged by two teacher say Mr. A and Ms. B. the teacher were each asked to rank order the eight essays from best to worst. The best essay was to be given a score of 1 and the worst a score of 8. Determine the degree of agreement between two teachers' ranking, given their rankings are as given below:

Rankings

Student	Teacher A (X)	Teacher (B)
John	6	6
Jill	4	7
James	7	5
Jane	5	3
Andrew	2	8
Jimmy	8	4
Joanna	3	2
Carolyn	1	1

Solution:

Let us present in the following table the information needed for calculating r:

X	Y	XY	X ²	Y ²
6	6	36	36	36
4	7	28	16	49
7	5	35	49	25
5	3	15	25	9
2	8	16	4	64
8	4	32	64	16
3	2	6	9	4
1	1	1	1	1
36	36	169	204	204

$$r = r_{xy} = r_{yx} = \frac{8(169) - (36)(36)}{\sqrt{[(8 \times 204 - 36^2)(8 \times 204 - 36^2)]}}$$

$$= \frac{1352 - 1296}{\sqrt{(336)(336)}} = \frac{56}{336} = 0.17$$

This weak correlation between the teachers' ranking suggests that they may have used different standards to rank the essays. Because both variables are ranked from 1 to n, one could have used Spearman's Rank Correlation Coefficient. But the answer will turn out to be exactly the same because there are no ties in the data.

Example 1.6: For each of the following calculate coefficient of determination from correlation coefficient and interpret both these statistics.

- (a) GRE and GPA: $r=.17$
- (b) Weight and GPA: $r=.10$
- (c) Number of safety-related failures during nuclear power plant standby and operation: $r=.68$
- (d) Attention/concentration ability and spatial ability: $r=-.34$

Solution:

- (a) There is a weak positive correlation between GRE and GPA; students with higher GRE scores tend to have higher GPAs. An $R^2 = 0.29$ means only about 3% of the variance in GPA is explained by GRE.
- (b) The correlation between weight and GPA is very weak and negative; heavier students tend to have lower GPAs. An $R^2 = .01$ means weight explains only about 1% of the variance in GPA.
- (c) The correlation between number of safety-related standby and operation failures is positive and moderate to strong. Plants with more standby failures tend also to have more operation failures. An $R^2 = .4624$ indicates that the number of standby failure explains about 46% of the variance of operation failures.
- (d) There is a weak to moderate negative correlation between attention/ concentration ability and spatial ability. Children with higher attention/ concentration ability tend to have lower spatial ability. An $R^2 = .1156$ means about 12% of the variance in spatial ability can be accounted for by attention/ concentration ability.

Example 1.7 Calculate Karl Pearson's coefficient of correlation for the data given below:

Independent Variable X: 3 7 5 4 6 8 2 7

Dependent Variable Y: 7 12 8 8 10 13 5 10

Solution: Let the assumed mean for x and y series be 5 and 9 respectively

Short-cut Method

The above direct method for calculating r is not convenient when

- (i) The terms of the series x and y are big and the calculation of x and y becomes difficult or
- (ii) The means x or y are not integers. In these cases we apply the following formula of assumed mean.

$$r_{xy} = \frac{\sum(dx dy) - \frac{(\sum dx)(\sum dy)}{n}}{\sqrt{\sum dx^2 - \frac{(\sum dx)^2}{n}} \sqrt{\sum dy^2 - \frac{(\sum dy)^2}{n}}}$$

Where $dx=x-a$, a is the assumed mean of x series, $dy = y-b$, b is the assumed mean of y series and n is number of observations of x and y .

Example 1.8. Calculate the coefficient of correlation for the following pairs of values of x and y .

$x:$ 17 19 21 26 20 28 26 27

$y:$ 23 27 25 26 27 25 30 33

Solution: Let the assumed means for x and y be 23 and 27 respectively. Then $dx=(x-23)$, $dy= (y-27)$. We have the following table.

Table: Computation of Correlation Coefficient

x	y	$dx=(x-23)$	$dy=(y-27)$	dx^2	dy^2	$dx dy$
17	23	-6	-4	36	16	24
19	27	-4	0	16	0	0
21	25	-2	-2	4	4	4
26	26	3	-1	9	1	-3
20	27	-3	0	9	0	0
28	25	5	-2	25	4	-10
26	30	3	3	9	9	9
27	33	4	6	16	36	24

Now

$$\begin{aligned}
 r_{xy} &= \frac{\sum(dx dy) - \frac{(\sum dx)(\sum dy)}{n}}{\sqrt{\sum dx^2 - \frac{(\sum dx)^2}{n}} \sqrt{\sum dy^2 - \frac{(\sum dy)^2}{n}}} \\
 &= \frac{[48 - 0]}{\sqrt{124 - 0} \sqrt{70 - 0}} \\
 &= \frac{48}{\sqrt{124 \times 700}} = \frac{48}{\sqrt{8680}} = \frac{48}{\sqrt{93.166}} = 0.515.
 \end{aligned}$$

Hence $r=0.515$.

Example 1.9: In two sets of variables X and Y with 50 observations each, the following data were observed.

$$\bar{x} = 10, S.D. \text{ of } x = 3, \bar{y} = 6, S.D. \text{ of } y = 2$$

Coefficient of correlation between X and Y is 0.3, However on subsequent verification it was found that one value of X(=10) and one value of Y(=6) were inaccurate and hence weeded out. With remaining 49 pairs of values, how is the original value of correlation coefficient affected.

Example -1.10 A computer while calculating the correlation coefficient between two variable X and Y obtained the following constants.

$$n = 30, \sum x = 120, \sum x^2 = 600, \sum y^2 = 250, \sum xy = 356.$$

It was however, later discovered at the time of checking that it had copied down two pairs of observations as: While the correct values were

X	Y	X	Y
8	10	8	12
12	7	10	8

Obtain the corrected values of the correlation coefficient between X and Y.

Solution: Here

$$\text{Correct } \sum x = 120 - 8 - 12 + 8 + 10 = 118$$

$$\text{Correct } \sum y = 90 - 10 - 7 + 12 + 8 = 93$$

$$\text{Correct } \sum x^2 = 600 - (8)^2 - (12)^2 + (8)^2 + (10)^2 = 600 - 64 - 144 + 64 + 100 = 556$$

$$\text{Correct } \sum y^2 = 250 - (10)^2 - (7)^2 + (12)^2 + (8)^2 = 250 - 100 - 49 + 144 + 64 = 309$$

$$\text{Correct } \sum xy = 356 - (8 \times 10) - (12 \times 7) + (8 \times 12) + (10 \times 8) = 356 - 80 - 84 + 96 + 80 = 368$$

Now,

$$\sum XY = 368; \bar{X} = \frac{118}{30} = 3.933; \bar{Y} = \frac{93}{30} = 3.1; \sum x^2 = 556, \sum y^2 = 309, N = 30,$$

$$\begin{aligned} \therefore r &= \frac{\frac{\sum XY}{N} - \bar{X}\bar{Y}}{\sqrt{\frac{\sum X^2}{N} - (\bar{X})^2} \times \sqrt{\frac{\sum Y^2}{N} - (\bar{Y})^2}} = \frac{\frac{368}{30} - (3.933)(3.1)}{\sqrt{\frac{556}{30} - (3.933)^2} \times \sqrt{\frac{309}{30} - (3.1)^2}} \\ &= \frac{12.27 - 12.19}{\sqrt{18.53 - 15.47} \times \sqrt{10.3 \times 9.61}} = \frac{0.08}{\sqrt{3.06} \times \sqrt{0.69}} = \frac{0.08}{1.453} = 0.055 \end{aligned}$$

You may try the following Exercises:

10.8 Exercises

E-1 An instructor believes that true-false tests are as effective as problem type tests in judging a student's proficiency in mathematics. A test consisting of half true false questions and half problem was given to ten students selected at random from a statistics class. The test results are as follows:

Student:	1	2	3	4	5	6	7	8	9	10
T-F type:	48	40	25	10	16	21	23	19	35	32
Problem type:	45	47	20	12	12	15	25	16	30	32

- Plot the data. Examine the scatter diagram and interpret it clearly.
- Calculate and interpret Pearson r to measure the strength of linear relation.
- Rank the scores on two variables, calculate Pearson r for ranked data. Which measure seems more appropriate, r or r_s ?

E-2 A sociologist working for the government of a large city collected data on the number of nonviolent crimes (in 1000s) reported and the increase (or decrease) of all crimes over the previous reporting period. Quarterly data are shown below:

Quarter	Nonviolent crimes	Increase (or decrease) in all crimes
1	7.2	14.1

2	6.4	14.5
3	6.6	13.3
4	7.3	13.6
5	7.5	15.2
6	6.9	15.7
7	7.1	15.3
8	7.4	14.8
9	7.6	16.1
10	7.3	16.6
11	7.1	16.2
12	7.0	15.9

- (a) Plot the nonviolent crime data versus quarter. Also plot the increase versus quarter on the same graph. Interpret each graph and summarize your impression. Does there appear to be a relationship between the two crime variables?
- (b) Compute and interpret the correlation coefficient and its square between nonviolent crimes and the increase in all crimes.

E-3 A researcher constructs an index of social status, X , (based upon education, income and occupation) and an index of liberal conservative political attitude, Y . The index of social status runs from 1 (low) to 5 (high) and the index of liberal conservative attitude from 1 (strong conservative) to 7 (strong liberal). The scores for 12 people on these indices are given below:

Person #:	1	2	3	4	5	6	7	8	9	10	11	12
X :	4	2	3	3	1	2	5	4	3	4	2	5
Y :	3	5	4	3	6	4	3	2	5	4	3	2

Construct a scatter diagram. Is there a linear trend among the data points? What is the direction of relationship? How would you describe this relationship?

E-4 Calculate the correlation coefficient between the height of sisters and height of brothers from the given data.

Height of sister (in cm)	64	65	66	67	68	69	70
Height of brother (in cm)	66	67	65	68	70	68	72

Hint: $\bar{x} = 67, \bar{y} = 68, \sum x = 0, \sum y = 0, \sum x^2 = 28, \sum y^2 = 34$. also $\sum xy = 25$

$$use\ r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}} = \frac{25}{\sqrt{28 \times 34}} = 0.81$$

E-5 Calculate the coefficient of correlation between x and y series from the following data:

	X series	Y series
Number of pairs of observation	15	15
Arithmetic mean	25	18
Standard Deviation	3.01	3.03
Sum of the squares of deviations from mean	136	138

Sum of the product of the deviations of x and y series from their respective means = 122.

Hint: $use\ r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}} = \frac{122}{\sqrt{136 \times 138}} = 0.89$

E-6 From the following data compute the coefficient of correlation between X and Y:

	X series	Y series
Arithmetic mean	15	28
Sum of the squares of deviations from arithmetic mean:	144	225

Sum of products of deviations of x and y series from their respective Means = 20, Number of pairs = 10.

E-7 The following table gives of husbands and wives at the time of their marriages. Calculate the correlation coefficient between the ages of husbands and wives.

Ages of husbands			Ages of wives	
	10-20	20-30	30-40	40-50
10-20	20	6	-	-
20-30	8	14	37	-
30-40	-	4	18	3
40-50	-	-	4	6

E-8 The coefficient of correlation between two variables X and Y is 0.64. Their covariance is 16. The variance of X is 9. What is the standard deviation of Y series.

E-9 Calculate the coefficient of correlation by concurrent deviation method.

Price:	1	4	3	5	5	8	10	10	11	15
Demand:	100	80	80	60	58	50	40	40	35	30

E-10 The following are the sources of 10 students and their IQ. In a class.

Students:	1	2	3	4	5	6	7	8	9	10
Scores:	35	40	25	55	85	90	65	55	45	50
I.Q. :	100	100	110	140	150	130	100	120	140	110

Find correlation between their score and I.Q.

10.9 Answers

E-4 - 0.81

E-5 - 0.89

E-6 - 0.11

E-7 - 0.7044

E-8 - 8.333

E-9 - 0.89

E-10 - 0.47.

10.10 Summary

Correlation means relationship and correlation coefficient is the measure of linear relationship between two variables X and Y. It is defined as

$$\frac{Cov(X,Y)}{\sqrt{var(X).Var(Y)}}$$

This is also known as product moment correlation coefficient or Karl Pearson's coefficient of correlation.

If the relationship X and Y is such that by increasing (or decreasing) one variable, other also increases (or decreases), i.e., change is in the same direction, correlation is said to be positive; otherwise negative.

When observations are taken on two variables X and Y over a group of individuals or units and these values are plotted on a graph by putting a point for a pair of observations of an individual or unit, the figure thus obtained is called Scatter or Dot Diagram. This diagram gives a rough idea about the relationship between the variables.

10.11 Further Readings

1. Goon , Gupta & Dasgupta : Fundamentals of Statistics Vol. I and Vol. II The World Press Pvt. Ltd., Kolkata.
2. Yule G.U. and Kerdall, M.G.: An Introduction to the Theory of Statistics, Charles Griffin and Company Ltd.
3. C.E. Weatherburn : Mathematical Statistics.

Unit-11 : Regression

Structure

11.1 Introduction

11.2 Objectives

11.3 Regressions

11.4 Purpose of Regression Analysis

11.5 Simple Linear Regression Model

11.6 Linear Regression and Least Square Principle

11.7 Exercises

11.8 Answers

11.9 Summary

11.10 Further Readings

11.1 Introduction

Prediction is a goal common to all sciences, including social and behavior sciences. Prediction is based on relationship between or among variables. It is the fact that crime rate and cost of living are correlated that enables us to predict crime rate from cost of living, or vice versa. Thus, the correlation between variables sets the stage for predicting one variable from another. The stronger the correlation between two variables, the more accurately can one be predicated from the other, and weaker the correlation, less accurate is the predictions. A referred to as bivariate regression analysis or simple linear regression. Regression analysis is a method of analyzing the variability of a dependent variable by resorting to information available on one or more independent variables. An answer is sought to the questions: What are the expected changes in Y as a result of changes in X?

11.2 Objectives

After going through this unit, you shall be able to

- Understand regression and obtain regression lines
- Use regression coefficients and their properties.

11.3 Regressions

Simple regression: The regression analysis confined to the study of only two variables at a time is called the simple regression.

Multiple Regression: The regression analysis for studying more than two variables at a time is known as multiple regression.

This chapter is confined only to the study of simple regression.

Linear Regression: If the regression curve is a straight line, then there is a linear regression between the variables under study. In other words, in linear regression the relationship between the two variables X and Y is linear.

Non linear Regression: If a curve or regression is not a straight line, i.e., not a first degree equation in the variables X and Y then it is called a non-linear or curvilinear regression. In this case the regression equation will have a functional relation between the variables X and Y involving terms in X and Y of the degree higher than one, i.e., involving terms of the type X^2 , Y^2 , X^3 , Y^3 , XY , etc.

Utility of Regression analysis:

1. The cause and effect relations are indicated from the study of regression analysis.
2. It establishes the rate of change in one variable in terms of the changes in another variable.
3. It is useful in economic analysis as regression equation can determine an increase in the cost of living index for a particular increase in general price level.
4. It helps in prediction and thus one can estimate the values of unknown quantities.
5. It helps in determining the coefficient of correlation as: $r = \sqrt{b_{yx} \times b_{xy}}$
6. It enable us to study the nature of relationship between the variables.
7. It can be useful to all natural. Social and physical sciences where the data are in functional relationship.

11.4 Purpose of Regression Analysis

An important use of statistical methods is to forecast or predict future events. Insurance companies sometimes set premiums on the basis of statistical predictions. The cost of automobile insurance for minors is greater than that for adults because age often correlates with frequencies of accidents. Colleges usually admit and reject applicants primarily on the basis of predictions about their probated future scholastic performance made from the scholastic aptitude tests and academic

performance made from the scholastic aptitude tests and academic performance high school. Delinquency and dropout prevention programs frequently use early indicators or predictors in identifying persons who appear likely to become delinquents or dropouts. In vocational counseling and personnel selection, implicit or explicit predictions of various job related criteria are made from variables such as age interests aptitudes, sex and experience. These examples involve prediction. The degree or reliance on statistical considerations in making these predictions varies greatly from one application to another. Insurance companies rely heavily on statistical predictions, whereas the selection of employees is rarely made on purely statistical considerations. By using statistical methods, the accuracy of predictions of a dependent variable (a criterion or outcome variable) from one or more independent (Predictor) variables can be maximized. In statistical terms, the dependent variable Y is said to be a function of the independent variable X. No causal relationship is assumed. Indeed, causation is beside the point in forecasting. The higher the greater the margin of error in predictions; the lower the correlation, the greater the margin of error in predictions. The simplest type of prediction involves predicting a dependent variable Y from only one independent variable X when both X and Y are normally distributed.

11.5 Simple Linear Regression Model

Suppose X and Y are perfectly linearly positively correlated, the simplest type of causal effect, the liner effect is represented by a straight line:

$$Y=a+bX.....(2.1)$$

Y is also known as dependents variable, response variable and endogenous variable, whereas X is known as in dependent variable, predictor variable exogenous variable. The line called the regression line of Y on X.

Where constant a= the value of Y when X=0.

This constant ‘a’ is also called the Y intercept because (when a line is plotted on a graph), ‘a’ is the value of Y at the point where the line crosses the Y-axis. We know that the equation of Y axis is X=0. b= the slope of the line.

The **slope** represents the change Y per unit increase in X. If X were a cause of Y, then a one unit increase in X would cause Y to change by b units. Thus b represents the effect of X on Y. If >0, Y is increasing with X. If b< 0, Y is decreasing with X.

The following figure shows a graph of a perfect positive linear relationship between X and Y where $a = 1$ and $b = 0.50$. Note that if there are only two points (a sample of size $n=2$), the relationship will always be perfect. The line is known if constants a and b are given. For example, $a = 1$, $b = 0.50$ gives the line as $Y = 1 + 0.50X$

We can plot this line on the graph paper, by taking only two points (i.e., A sample of size $n=2$), The relationship will be perfect.

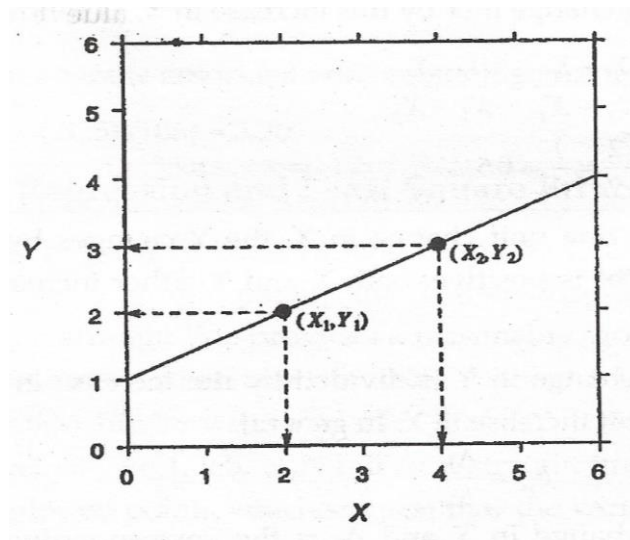


Fig.2.1

The line can be plotted by first taking any two values of X and computing their corresponding values of Y from the formula. The two pairs of X, Y values form two pairs of coordinates (X,Y) through which the line must pass. Suppose our sample on X variable yields. The coordinates in the above figure are determined by arbitrarily choosing $X_1=2$ and $X_2=4$.

If $X_1 = 2$, $Y_1 = 1 + 0.50(2) = 2$ giving the point $(X_1, Y_1) = (2, 2)$

If $X_2 = 4$, $Y_2 = 1 + 0.50(4) = 3$ giving the point $(X_2, Y_2) = (4, 3)$

The plotted line $Y = 1 + 0.50 X$ passes through these points (2,2), (4,3). This line can be used for further prediction of Y value for any X value.

Conversely, if we were given the straight line drawn on the graph but did not know equation of the line we could determine it from the graph.

The slope 'b' could be determined by first selecting any two points on the line. One of these points could be the point at which the line crosses the Y-axis which is (0,a). In the above figure it is (0,1) from each of the two points drop a vertical line down to the X-axis to determine the X-values and run a horizontal line to the Y axis to determine Y values. The value of the slope will then be

computed by subtracting the smaller X value from the larger X value to get the increase in X and then dividing the corresponding change in Y by this increase in X, as follows:

$$b_{yx} = \frac{Y_2 - Y_1}{X_2 - X_1} = \frac{Y_1 - Y_2}{X_1 - X_2}$$

Here

$$b = \frac{3 - 2}{4 - 2} = \frac{1}{2} = 0.5$$

Thus with one unit change in X and the Y changes by .50 unit, and since the sign of 'b' is positive, both X and Y either increase or decrease simultaneously.

When the change in Y is divided by the increase in X we get the change in Y per unit increase in X. In general

$$b_{yx} = \frac{\Delta y}{\Delta x} \dots \dots \dots (2.2)$$

Where Δy is change in Y and Δx is the corresponding change in X values.

Example 2.1: Suppose you are told that for a group of 20 students, there is a perfect linear relationship between grade point average (Y) and scores on and intelligence test (X). Suppose you are also told that the equation describing the relationship is: $Y=1.00+0.025(X)$

If an individual obtained a score of 100 on the inelegance test score of 97? And of 108?

Solution: Here $X = 100$

The grade point average associated with an intelligence test score to 100 is

$$Y=1.00+(0.25)(100)=1.00+2.50=3.50$$

For $X=97$: The grade point average associated with an intelligence test score of 97 is

$$Y= 1.00+(0.25) (97)= 3.425$$

For $X = 108$: The grade point average associated with intelligence test score of 108 is

$$Y= 1.00+(0.25) (108)= 3.70.$$

Check your progress

1. Define the Regression Analysis.
2. What do you mean by Simple Linear Regression Model?

11.6 Linear Regression and Least Square Principle

Suppose N pair of observations (x_i, y_i) , $i = 1, 2, 3, \dots, N$ are given. When plotted on a graph paper the scatter diagram may give an impression that a straight line provides a reasonably good fit to the data. It means that one may be justified in using a linear regression model to explain the relationship between variables X and Y . It does not mean that all points (X_i, Y_i) , $i = 1, 2, 3, \dots, N$ fall on a straight line. It is overall pattern of the plotted points which suggest that the variables X and Y are related by the relation of the type either $Y = a + bx$ or $X = \alpha + \beta Y$, where (a, b) or (α, β) are the unknown quantities. The line of regression is the straight line which gives the best fit in the least square sense to the given bi-variate frequency or probability distribution.

Once it is decided that in a given problem the regression is approximately linear, say $Y = a + bX$, then the problem is to use the data, to obtain estimates $\hat{a}$ and $\hat{b}$ of a and b respectively such that the estimated regression line $Y = \hat{a} + \hat{b}X$ is in least square sense provides the best possible fit to the given data. Let us understand what does it mean

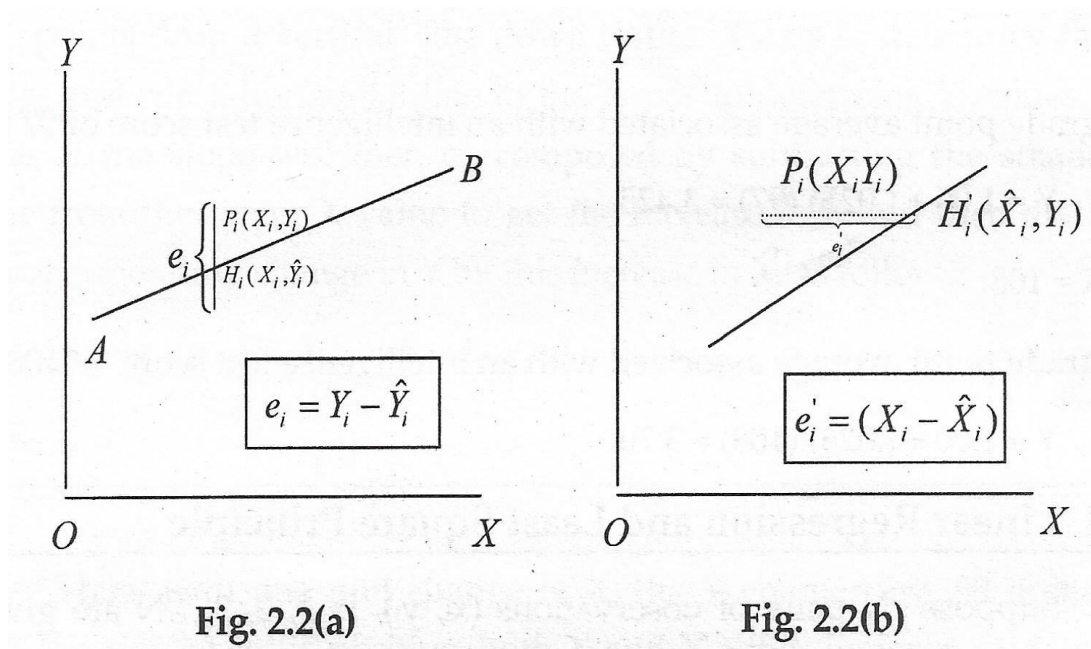
Let $Y = a + bx \dots \dots \dots (2.3)$

Be the regression line.

y_i = actual value of Y for $X = x_i$; $\hat{Y}_i$ = estimate value of Y for $X = x_i$

$e_i = Y_i - \hat{Y}_i$ = Error or Residual in Y for $X = x_i \dots \dots \dots (2.4)$

The error e_i is the error in prediction using (2.3).



In the Fig 2.2 (a) the residuals are the vertical distance between each point and the line while in Fig 2.2(b), the residuals are the horizontal distance between each point and the line.

A logical way of evaluating the predictive power of the equation is to compute the sum of all the residuals. But since some residuals are positive, some negative, and some zero when they are all summed, we get a sum of zero. Therefore, we squared residual that is always positive.

$$e_i^2 = (Y_i - \hat{Y}_i)^2 \dots \dots \dots (2.5)$$

After squaring each residual sum them all to get a sum of squared residuals or error sum of squares (Error S.S.)

$$\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2 \dots \dots \dots (2.6)$$

The principle of least square requires that 'a' and 'b' be so chosen that the error sum of square $\sum e_i^2$ is minimum. The estimates $\hat{a}$ and $\hat{b}$ are known as least square estimates of a and b respectively $Y = \hat{a} + \hat{b}X$ and as regression line of Y on X.

A natural question would be whether there is another line that could reduce this total amount of squared error $\sum e_i^2$ i.e. is there a line that would make even better predictions? In other words whether there exists a line for which the error SS is minimum? No, There does not exist such a line. This line $Y = \hat{a} + \hat{b}X$ is the best fit line. This line minimizes the sum of squared residuals and is the line of best fit. Here, the residual SS for ordinate Y_i 's are minimized. Hence the line is said to be regression line of Y on X. It is of the form $\bar{Y} = b_{yx} (X - \bar{X})$ or $Y = \hat{a} + \hat{b}X$. This question suggests that minimizing the sum of squared values of residuals might be a good criterion for selecting the line that fits the data best. The line that minimizes the squared residuals is the line of best fit. The line that minimizes the squared residuals is called the **least-squares line**.

The equation that describes this line is referred to as the least-squares equation. The mathematical theory assures that the least squares line (also called the regression line or the prediction line) fits better than any other line would fit. Therefore we use the least squares line or more specifically, the least squares slope, as the best estimate of the linear effect of X on Y.

Similarly, if a straight line is so chosen that the sum of square of errors or residuals $\sum e_i'^2$, where, $e_i' = X_i - \hat{X}_i$ parallel to the x-axis is minimized, then the resulting straight line $X = \hat{\alpha} + \hat{\beta}Y$ is the regression line of X and Y, it gives the best estimate of X for a given value of Y.

Thus in general there are two regression lines namely the regression line of Y on X and regression line of X on Y.

The regression lines 'Y on X' and 'X on Y' are obtained by using the principles of least squares. It means the constants (a, b) of regression line Y on X and (α , β) of the regression line X on Y are determined by using the principle of least squares.

11.7 Exercises

E-1 comment on the following results obtained from given data:

- (i) The two regression coefficient are 1.5 and 0.6 and the coefficient of correlation is 0.90.
- (ii) The two regression coefficient are -2.7 and -0.3 and the coefficient of correlation is 0.90.
- (iii) The two regression coefficient are greater than unity.
- (iv) Coefficient of regression of Y on X is 4.2; and coefficient of regression of X on Y is 0.5.
- (v) The two regression obtained by a student are 2.58 and 0.48.
- (vi) The coefficient of correlation between X and Y is 0.85 and one of the regression coefficient is -0.21.
- (vii) "If r is correlation coefficient then r^2 is the proportion of total variation in the dependent variable which is explained by regression analysis."

E-2 Indicate whether each of the following statements are true or False:

- (i) The product of the regression coefficient obtained from a bivariate data is always less than unity.
- (ii) The lines of regression do not exist if correlation between the variables is zero.
- (iii) The lines of regression are based on the principle of least squares.
- (iv) The line of regression of Y on X can be used to estimate the average value of Y for a given value of X.
- (v) The two regression lines are mutually perpendicular if X and Y are independent.
- (vi) The two lines of regression are perpendicular to each other if $r = \pm 1$.
- (vii) Regression coefficient are independent of change of origin.

E-3 Fill in the blanks

- (i) The two regression coefficient are ofsign.
- (ii) The understood of two..... coefficients gives us the value of correlation coefficient.
- (iii) The variable we are trying to predict is called the

- (iv) Both the regression coefficients cannotunity.
- (v) If a regression coefficient is negative then the correlation between the two variables would also be
- (vi) The coefficient of determination is a real number lying between and
- (vii) The regression analysis helps us to study the relationship, between the variables.
- (viii) If both the regression coefficients are negative the correlation coefficient would be.....
- (ix) Regression analysis is used to study..... between the variables.
- (x) If $r = \pm 1$ the two regression lines areto each other.
- (xi) If $r = 0$, the two regression lines are to each other.

11.8 Solutions/ Answers

E-1 i) wrong ii) wrong iii) wrong iv) wrong v) wrong vi) wrong vii) ?

E-2 (i) True (ii) False (iii) True (iv) False (v) False (vi) False (vii) False (viii) True (ix) True (x) False (xi) False (xii) True (xiii) True (xiv) True (xv) True (xvi) True (xvii) True (xviii) False (xix) False (xx) False.

E-3 (i) same (ii) Regression (iii) Dependent variable (iv) Exceed (v) Negative (vi) 0 and 1 (vii) Nature (viii) Negative (ix) Dependent (x) Parallel (xi) Perpendicular.

11.9 Summary

Regression is mainly concerned with bringing out the nature of relationship between variables and using it to know the best approximate value of one variable corresponding to a known value of the other variable.

11.10 Further Readings

1. Goon , Gupta & Dasgupta : Fundamentals of Statistics Vol. I and Vol. II The World Press Pvt. Ltd., Kolkata.

2. Yule G.U. and Kerdall, M.G.: An Introduction to the Theory of Statistics, Charles Griffin and Company Ltd.
3. C.E. Weatherburn : Mathematical Statistics.

Unit-12: Line of Regression

Structure

12.1 Introduction

12.2 Objectives

12.3 Regression Line

12.4 Some important properties relating to regression coefficient

12.5 Exercises

12.6 Answers

12.7 Summary

12.8 Further Readings

12.1 Introduction

Bivariate regression refers to the case in which only one X and one y are being analyzed at a time. It is through bivariate regression analysis that the correlation between X and Y is used in predicting one variable from the other variable, called the criterion variable or dependent variable or Y variable, from the other variable called the predictor variable or independent variable or X variable.

One way of facilitating predictions is to obtain a simple linear equation that fits, or represents the available data. This equation can then be used to study how a change in X variable relates to a change in Y variable. In this chapter we discuss simple linear regression that helps us find such a prediction equation.

12.2 Objectives

After going through this unit you shall be able to

- Understand regression line
- Use some important properties relating to regression coefficients

12.3 Regression Lines

Suppose that there are N pairs of observations (X_i, Y_i) , $i = 1, 2, 3, \dots, N$. Let the regression line of Y on X be

$$Y = a + bX$$

So that error sum of squares $= \sum e_i = \sum (Y_i - \hat{a} - b\bar{X}_i)^2 = S^2 \dots \dots \dots (2.7)$

The desired estimates of a and b are obtained by solving the simultaneous equations, called the normal equations.

$$\frac{\delta S^2}{\delta a} = 0, \frac{\delta S^2}{\delta b} = 0$$

or

$$\sum_i (Y_i - a - bX_i) = 0$$

$$\sum_i X_i (Y_i - a - bX_i) = 0$$

$$\text{and i. e., } \sum_i Y_i = Na + b \sum_i X_i$$

$$\text{and } \sum_i X_i Y_i = a \sum_i X_i + b \sum_i X_i^2 \dots \dots \dots (2.8)$$

The roots of the equations are

$$b = \frac{N \sum_i X_i Y_i - (\sum_i X_i)(\sum_i Y_i)}{N \sum_i X_i^2 - (\sum_i X_i)^2} \dots \dots \dots (2.9)$$

$$= \frac{\sum_i \frac{X_i Y_i}{N} - \bar{X}\bar{Y}}{\frac{(\sum_i X_i)^2}{N} - \bar{X}^2}$$

$$\frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})/N}{\sum_i (X_i - \bar{X})^2/N}$$

$$= \frac{Cov(X, Y)}{var(X)} = r \frac{\sigma_y}{\sigma_x} \dots \dots \dots (2.10)$$

$$\text{And } \hat{a} = \bar{Y} - \hat{b}\bar{X} \dots \dots \dots (2.11)$$

Substituting these value in (2.3) we have the desired prediction formula:

$$Y = \bar{Y} + r \frac{\sigma_y}{\sigma_x} (X - \bar{X}) \dots \dots \dots (2.12)$$

The line given by Eqn. (2.12) is known as regression line of Y and X.

The coefficient b is the amount by which the predicated value Y increases for a unit increment in the value of X . It is called the regression coefficient of Y on X . It is also written as b_{yx} .

The regression line of Y on X is used to get the best estimate of variable Y for any specified value of X .

The regression line of Y on X is also written as

$$Y - \bar{Y} = b_{yx} (X - \bar{X})$$

or

$$Y = \bar{Y} + r \frac{\sigma_y}{\sigma_x} (X - \bar{X})$$

Similarly if we are interested in predicting X from Y , we use the regression line of X on Y , which has the equation.

$$X = \bar{X} + r \frac{\sigma_x}{\sigma_y} (Y - \bar{Y}) \dots \dots \dots (2.13)$$

$r \frac{\sigma_x}{\sigma_y}$ is the amount by which the predicated value X increases for a unit increment in Y , is the regression coefficient of X on Y . It is denoted by b_{xy} .

It may be noted that both the regression lines pass through the point which is their point $(\bar{X}, \bar{Y})$ of intersection.

We usually designate one variable as dependent (Y) and the other as independent (X) when using regression to estimate the effect of X on Y . It is also possible to calculate a regression equation that uses the variable labeled Y as predictor of the variable labeled X . Two researchers, for instance, might disagree about the direction of effect.

It may be instructive to examine how regression coefficients differ in the two situations.

The regression equations or regression X on Y is written as:

$$X' = a_{xy} + b_{xy}(Y)$$

Where

$$b_{xy} = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (Y_i - \bar{Y})^2} = \frac{n \sum_i X_i Y_i - (\sum_i X_i)(\sum_i Y_i)}{n \sum_i Y_i^2 - (\sum_i Y_i)^2} \dots \dots \dots (2.14)$$

$$\text{and } a_{xy} = \bar{X} - b_{xy}(\bar{Y})$$

Thus intercept and slope are asymmetric statistics. If the variances of X and Y were equal, then $b_{xy} = b_{yx}$.

The fact that there are two different regression lines one when Y is treated as the dependent variable and another when X is treated as dependent, does not mean they can both be used as measures of the effects of each variable on the other. They cannot both be correct. We have to choose one variable as dependent and the other as independent and then use the appropriate formulas.

Lines of Regressions:

A line of regression is the line which gives the best estimate of one variable for any given value of the other variable.

I. Line of regression of X on Y. It is the line which gives the best estimate for the values of X for a specified value of Y.

It is given by, $X = \bar{X} + r \frac{\sigma_y}{\sigma_x} (Y - \bar{Y})$

Where $\bar{X}$, $\bar{Y}$ are means of X series and Y series respectively are S.D. of X and Y series respectively and r is the correlation coefficient between X and Y.

It can also be put in form: $X = \hat{\alpha} + \hat{\beta}_y$

Where $\hat{\alpha}$ is intercept of the line and $\hat{\beta}$ is the slope of the line X on Y.

II. Line of Regression of Y on X. It is the line which gives the best estimates for the values of Y for any specific values of X.

Regression equation of Y on X is given by:

$$Y = \bar{Y} + r \frac{\sigma_y}{\sigma_x} (X - \bar{X})$$

It can also be put in the form $Y = a + bX$,

Where a is the intercept of the line and b is the slope of the line Y on X.

Regression Coefficients

The regression coefficient of Y on X is $b_{yz} = r \frac{\sigma_y}{\sigma_x}$ and that of x on y is $b_{xy} = r \frac{\sigma_y}{\sigma_x}$.

Interpretation of a and b

(i) Interpretation of intercept: By definition intercept a is the predicated value of Y when X=0. There are several factors that affect interpretation of a. While interpreting intercept, pay attention to the following:

- (a) **The measurement of level of X:** If X is a ratio variable with a true zero value the intercept may represent the value of Y that is caused by the absence of whatever X is measuring. For example, the absence of income over some period of time may cause a high level of alienation. But if X is not a ratio level variable, we cannot give the same interpretation to a because $X=0$ does not mean the absence of what X is measuring because the zero on the scale is not a true zero.
- (b) **Range of X:** One must pay attention to when interpreting intercept is whether there are any cases that take on the value $X=0$. For example, if X is age and you were investigating its effect on earnings of employed persons there would not be any persons in the study who were zero years old, or even close to it .So, it will not be justifiable and meaningful to extrapolate predictions beyond the range of the observed values of X.
- (c) **Distribution of X:** There may not be enough cases at $X=0$ to use the intercept as a prediction of Y. For example, if X is the numbers of years of schooling there will be very few individuals with a score of 0. Since there are very few cases at this value of X, they will not have much weight in determining the slope of the line and hence prediction may be very bad near zero.

Because of these reasons there is little or no interest in the intercept we are primarily interested in the slope of the regression equation. We focus on how much and in what direction Y changes as a result of a change in X, that is the slope of the line.

- (III) **Interpretation of Slope:** Slope is the change in Y per unit change in X. To be more accurate for cross sectional data. It is a theoretical assumption to interpret the slope change in Y per unit X. We have not measured any changes in either X or the Y. So, to represent what has been empirically observed more accurately, we should say that the slope is the difference in Y per unit difference in X. Also not all variables that are causes of other variables produce their effects through changes. Race and gender do not change, yet many things effected by these variables. Similarly, as another example although the number of siblings (X) that a child has many change as his or her parents have more children, these changes are usually completed before the child begins having children (Y). Similarly although education changes as a person progresses through higher and higher levels of schooling they typically are not working at jobs where their earnings are affected by these

changes until after they have completed their education. Thus in all such situation it is best to interpret the regression slope as the differences in Y for a positive unit difference in X.

Example 21: In a sample of 143 college students the regression equation for prediction weight in pounds (Y) from height in inches (X) was determined to be $Y = 238 + 5.6X$.

- (a) Explain what the slope in this equation tells us
- (b) What is the literal meaning of the intercept in this equation? Why should we not expect a meaningful intercept in an equation for predicting college students' weights from their heights/

Solution:

- (a) Slope $b = 5.6$ means for every inch increase in height average or predicted weight increases by 5.6 pounds.
- (b) Intercept $a = -238$ means a college student 0 inches tall is predicted to weigh -238 pounds. This makes no sense because height of 0 inches lies for outside of the normal range of this variable –no college student is 0 inches tall.

Example 2.2: Following are the data on nicotine concentrations in cabin air (smoking sections) for 26 commercial airline flights. These data were collected to study the effects of seating segregation (smoking or nonsmoking) on air quality. Nicotine concentrations are measured in micrograms per cubic meter of air. Two other variables reported are the number of passengers in the section and the estimated number of cigarettes smoked. Values of these variables are missing for six flights.

Flights	Number of passengers in smoking section	Number of cigarettes smoked	Nicotine concentration
1	13	26	.03
2	-	-	.08
3	25	88	.4
4	20	37	.6
5	21	50	.07
6	22	37	.07
7	-	-	2.1
8	-	-	2.3
9	10	17	3.1

10	-	-	4.5
11	24	20	8.6
12	10	17	8.8
13	10	23	10.2
14	17	32	10.5
15	-	-	11.0
16	-	-	14.9
17	35	123	18.7
18	11	6	22.1
19	7	11	30.2
20	15	19	39.5
21	20	30	45.0
22	22	30	45.0
23	20	17	57.1
24	22	84	59.8
25	23	38	76.7
26	23	31	112.4

The regression of nicotine concentration (Y) on number of passengers in smoking section (X) is describe by the regression equation $Y = 9.07 + .99(X)$

- Interpret the value of the intercept
- Interpret the value of the intercept
- Construct a scatter plot of the data and draw in the regression line. Indicate the graphical meanings of intercept and slope.

Indicate the graphical meanings of intercept and slope

Solution:

- Slope $b = 0.99$ means for every additional smoking section passenger, predicted nicotine concentration increases by 0.99 ug/m^3 of air.
- Intercept $a = 9.07$ means even if there are no passengers in the smoking section, the nicotine concentration is predicted to be 9.07 ug/m^3 .

Example 2.3: The regression equation describing the relationship between cumulative grade point average (Y) and the average number of classes missed per month (X) in a sample of college students is: $Y = 3.04 - .064(X)$.

- (a) What is the predicted GPA for a student who never misses class?
- (b) What is the predicted GPA for a student who misses ten classes per month?

Solution: (a) if $X_i = 0, Y = 3.04 - .064(0) = 3.04$

(b) if $X_i = 10, Y = 3.04 - .064(10) = 2.04$

Check your progress

1. Define the Regression lines.
2. What do you mean by interpretation of intercept?

12.4 Some Important Properties Relating to Regression Coefficient

Consider any one of the regression lines, say that of Y on x, It has the following properties:

Property I.

let $u = \frac{x - a}{c}$ and $v = \frac{y - B}{d}$, where $c > 0$ and $d > 0$ (2.15)

Then the regression coefficient of y on x denoted by b_{yx} for the sake of definiteness is

$$b_{yx} = \frac{\text{cov}(x, y)}{\text{var}(x)} = \frac{cd \text{cov}(u, v)}{c^2 \text{var}(u)} = \frac{d}{c} \times \frac{\text{cov}(u, v)}{\text{var}(u)} = \frac{d}{c} \times b_{vu} \dots \dots (2.16)$$

Or

$$b_{yx} = \frac{d}{c} \times \frac{n \sum_i u_i v_i - (\sum_i u_i)(\sum_i v_i)}{n \sum_i u_i^2 - (\sum_i u_i)^2} \dots \dots \dots (2.17)$$

The other constants in the regression equation are, in terms of u and v.

$$\bar{y} = B + d\bar{v} \text{ and } \bar{x} = A + c\bar{u}.$$

Property II.

$$\text{Since } \hat{y} = \bar{y} + r \frac{\sigma_y}{\sigma_x} (x_i - \bar{x}), \dots \dots \dots (2.18)$$

$$\therefore \hat{Y} = \frac{1}{n} \sum_i \hat{Y}_i = \frac{\left[n\bar{y} + r \frac{\sigma_y}{\sigma_x} (x_i - \bar{x}) \right]}{n} \dots \dots \dots (2.19)$$

Dividing both sides by n and remembering that $\sum_i (x_i - \bar{x}) = 0$, we have

$$\hat{\bar{Y}} = \bar{y} \dots \dots \dots (2.20)$$

In words the mean of the observed values of y is equal to the mean of the corresponding predicted values..

Property II the mean of the errors of estimates, $e_i = y_i - \hat{Y}_i$, is zero Since

$$e = \frac{1}{n} \sum_{i=1}^n e_i = \frac{1}{n} (y_i - \hat{Y}_i) = \bar{y} - \hat{\bar{Y}} = 0$$

Property III.

Again the residual variance, var (e), is given by

$$\begin{aligned} n \text{ var } (e) &= \sum_i e_i^2 \quad [\bar{e} = 0] \\ &= \sum_i (y_i - \hat{Y}_i)^2 \dots \dots \dots (2.21) \\ &= \sum_i \left\{ (y_i - \bar{y}) - r \frac{\sigma_y}{\sigma_x} (x_i - \bar{x}) \right\}^2 \\ &= \sum_i (y_i - \bar{y})^2 - 2r \frac{\sigma_y}{\sigma_x} \times \sum_i (x_i - \bar{x})(y_i - \bar{y}) + r^2 \frac{\sigma_y}{\sigma_x} (x_i - \bar{x})^2 \\ &= n s_y^2 - 2r \frac{\sigma_y}{\sigma_x} \times n r \sigma_x \sigma_y + r^2 \frac{\sigma_y}{\sigma_x} \times n \sigma_x^2 \dots \dots \dots (2.22) \end{aligned}$$

$$\text{Hence, var } (e) = \sigma_x^2 (1 - r^2) \dots \dots \dots (2.23)$$

The standard deviation of e, which is called the standard error of estimate of y its linear regression on X is denoted by We have, then,

$$\sigma_{yx} = \sigma_y \sqrt{1 - r^2} \dots \dots \dots (2.24)$$

Since var (e) ≥ 0, we have

$$r^2 \leq 1 \quad \text{or} \quad -1 \leq r \leq 1 \dots \dots \dots (2.25)$$

A result which has already been proved in a different way.

Property IV.

We have seen that

$$b_{yx} = r \frac{\sigma_y}{\sigma_x} \dots \dots \dots (2.26)$$

and

$$b_{xy} = r \frac{\sigma_x}{\sigma_y} \dots \dots \dots (2.27)$$

Hence

$$b_{yx} \times b_{xy} = r^2$$

or

$$/r/ = \sqrt{b_{yx} \times b_{xy}} \dots \dots \dots (2.28)$$

Thus numerically the correlation coefficient is the geometric mean of the two regression coefficients. As regards the sign of r, it is the same as the common sign of the two regression coefficients.

Property V.

If one of the regression coefficient is greater than unity then the other is less than unity.

Proof. Let $b_{yx} > 1$. Also we know that

$$\begin{aligned} r^2 &\leq 1 \text{ and } r^2 = b_{yx} \times b_{xy} \\ \rightarrow b_{yx} \times b_{xy} &\leq 1 \quad [r \leq 1] \\ \rightarrow b_{yx} &\leq \frac{1}{b_{yx}} < 1. \end{aligned}$$

Property VI.

Arithmetic mean of the regression coefficient is greater than the correlation coefficient.

Property VII.

Regression coefficients are Independent of change of origin but not of scale.

Property VIII.

Both regression coefficient are independent of change of origin but not of scale.

Property IX.

The sign of correlation is same as that of regression coefficients, i.e., r.0 if $b_{xy} > 0$; and r<0 if regression coefficients are negative.

$$b_{yx} = r \frac{\sigma_y}{\sigma_x} = \frac{r \sigma_y \sigma_x}{\sigma_x} = \frac{Cov(X, Y)}{\sigma_x^2}$$

Similarly

$$b_{xy} = \frac{Cov(X, Y)}{\sigma_x^2}$$

And

$$r = \frac{Cov(X, Y)}{\sqrt{\sigma_x \sigma_y}}$$

The sign b_{yx} , b_{xy} of and r is same as $Cov. (X, Y)$

Example 2.4: Using the data in Example 12.11 the statistic on median education (X) and average teacher salary (Y) are given below:

Mean of X = 12.47 Mean of Y = 19991.5

Sum of squares for X Sum of squares for Y

X = .5629 Y = 197774971.6

Sum of cross products = 5110.49

- Use these statistics to find the regression equation for predicting average teacher salary from a state's median education level.
- Find the correlation coefficient and coefficient of determination and describe the relationship.
- In what state is the average teacher salary the farthest below what we would expect or predict, based on median education? In what state is the average teacher salary the farthest above what would predict?

Solution:

(a) $b = 5110.49 / .5629 = 9078.86$

$a = 19991.5 - 9078.86 (12.47) = 93221.9$

$Y = -93221.9 + 9078.86 (X)$

(b) $r = 5110.49 / (\sqrt{.5629} \sqrt{197774971.6}) = .484$

$r^2 = R^2 = .484^2 = .235$

There is a moderate positive relationship between both variables. States with higher median education levels tend also to have higher average teacher salaries. Median education level explains 23.5% of the variation in average teacher salaries.

- In Vermont the average teacher salary is relatively low (16299). Based on Vermont's average median education level (912.6), we would predict a much higher average

salary: $Y = 21172$. Vermont has the largest negative residual (-4873) among the 21 states. Similarly, New York is found to have the largest residual.

Regression Fallacy

Unless $r=1$ or -1 all predictions of Y from X involve a regression towards the mean. Francis Galton first documented this regression effect in studying the relationship between the characteristics of parents and their children.

In virtually all test-retest situations, the bottom group on the first test will on an average show some improvement on the second test and the top group will on average fall back. This is called the regression effect. The regression fallacy consists in thinking that the regression effect must be due to something important not just the spread around a line.

Example 2.5: For example, a preschool program attempts to boost children's IQs. The children are tested when they enter the program. (the pre-test), and again when they leave (post-test). On both occasions, the scores average out to nearly 100, with a standard deviation of about 15. The Programme seems to have no effect. A closer look at the data, however, seems to show something very surprising. The children who were below average on the pre-test show an average gain of about 5 IQ point at the post-test. Conversely those children who were above average on the pre-test. An average loss of about 5 IQ points at the post test. What does this prove? Does the program operate to equalize intelligence? Perhaps when the brighter children play with the duller ones, the difference between the two groups tends to be diminished is this good or bad?

Solution: These speculations may be very interesting, but the fact is that nothing interesting is going on, good or bad. The children cannot be expected to score exactly the same on two tests. There will be differences between the two scores. Nobody would think these differences mattered, or needed any explanation. But these differences make the scatter diagram for the test scores spread out around the standard deviation line into the football shaped cloud. It is just this spread around the line which makes the bottom group come-up and the top group come down. There is nothing else to it.

Some More Solved Examples are given below.

Example 2.6. Find both the regression equations from the following data:

$$\begin{array}{lll} \sum X = 60 & \sum Y = 40 & \sum XY = 1150 \\ \sum X^2 = 4160 & \sum Y^2 = 1720 & N = 10 \end{array}$$

Solution. The regression coefficient b_{xy} and b_{yx} given by:

$$b_{xy} = \frac{\sum XY - \frac{\sum X - \sum Y}{N}}{\sum Y^2 - \frac{(\sum Y)^2}{N}} = \frac{1150 - \frac{60 \times 40}{10}}{1720 - \frac{(40)^2}{10}} = \frac{910}{1560} = 0.58$$

and

$$b_{yx} = \frac{\sum XY - \frac{\sum X - \sum Y}{N}}{\sum X^2 - \frac{(\sum X)^2}{N}} = \frac{910}{4160 - \frac{(60)^2}{10}} = 0.24$$

Also

$$\bar{X} = \frac{\sum X}{N} = \frac{60}{100} = 6; \quad \bar{Y} = \frac{\sum Y}{N} = \frac{40}{100} = 4$$

Now, the regression equation of Y on X is:

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

$$Y - 4 = 0.24(X - 6) \text{ or } Y = 0.24X + 2.56.$$

The regression of X on Y is:

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

$$X - 6 = 0.58(Y - 4) \text{ or } X = 0.58Y + 3.68.$$

Example 2.7. In a partially destroyed record the following data are available.

Variance of x= 25.

Regression equation of x on y : $5x - y = 22$.

Regression equation of y on x : $64x - 45y = 24$.

Find (a) mean values of x and y; (b) coefficient of correlation between x and y.

(c) Standard deviation of y.

Solution. (a) The mean values of x and y lie on the regression lines and are obtained by solving the given regression equations.

$$5\bar{x} - \bar{y} = 22 \dots \dots \dots (1)$$

$$\text{and } 64\bar{x} - 45\bar{y} = 24 \dots \dots \dots (2)$$

Multiplying the equation (1) by 45. We get.

$$225\bar{x} - 45\bar{y} = 990 \dots \dots \dots (3)$$

Subtracting (2) from (3) we get $161\bar{x} = 966 \Rightarrow \bar{x} = 6$.

Putting $\bar{x} = 6$ in (i) we get $30 - \bar{y} = 22 \Rightarrow \bar{y} = 8$.

Hence $\bar{x} = 6$ and $\bar{y} = 8$.

(b) The regression equation of y on x is:

$$64x - 45y = 24 \text{ or } y = \frac{64}{45}x - \frac{24}{45} \rightarrow y = -\frac{8}{15} + \frac{64}{45}x$$

$$b_{yx} = \frac{64}{45}$$

Again regression equation of x on y is : $5x - y = 22$ or $x = \frac{22}{5} + \frac{1}{5}y$

$$\therefore b_{xy} = \frac{1}{5}$$

But

$$r = \pm \sqrt{b_{xy} \times b_{yx}} = \sqrt{\frac{64}{45} \times \frac{1}{5}} = \frac{8}{15}$$

(+ve sign with r is taken as both the regression coefficient b_{xy} and b_{yx} are positive).

Hence, the coefficient of correlation $r = 8/15$.

(c) Now it is given that variance of x = $\sigma_x^2 = 25 \Rightarrow \sigma_x = 5$

Also

$$r = \frac{8}{15}, \quad b_{yx} = \frac{64}{45}$$

But we know that

$$b_{yx} = r \frac{\sigma_y}{\sigma_x} \Rightarrow \frac{64}{45} = \frac{8}{15} \times \frac{\sigma_y}{5} \Rightarrow \sigma_y = \frac{40}{3} = 13.33$$

Hence the standard deviation of y = 13.33.

Example 2.8: From the following results, obtain the two regression equations and estimate the yield when the rainfall is 29 cms. And the rainfall, when the yield is 600 kg:

	Yield in kg.	Rainfall in cms.
Mean	508.4	26.7
S.D.	36.8	4.6

Coefficient of correlation between yield and rainfall is + 0.52.

Solution: Let x represent rainfall and y represent yield.

Regression of y on x is given by $y - \bar{y} = r \frac{\sigma_y}{\sigma_x}(x - \bar{x})$

$$\text{or } y - 508.4 = 0.52 \times \frac{36.8}{4.6}(x - 26.7) \Rightarrow y - 508.4 = 0.52 \times 8(x - 26.7)$$

$$\Rightarrow y - 508.4 = 4.16x - 111.072 \Rightarrow y = 4.16x + 397.328$$

When $x=29$, yield $y=4.16 \times 29 + 397.328 = 517.968 \text{ kg}$.

Regression of x on y is given by $x - \bar{x} = r \frac{\sigma_y}{\sigma_x} (y - \bar{y})$

$$\text{or } x - 26.7 = 0.52 \times \frac{4.6}{36.8} (y - 508.4) = 0.065(y - 508.4) = 0.065y - 33.046$$

$$\text{or } x = 0.065y - 6.346$$

When $y=600$, rainfall $x=0.065 \times 600 - 6.34 = 32.654 \text{ cms}$.

Example 2.9 For a bivariate data the mean value of X is 20 and the mean value of y is 45. The regression coefficient of Y on X is 4 and that of X on Y is $(1/9)$. Find

- (i) The coefficient of correlation
- (ii) The standard deviation of X if the standard deviation of Y is 12.
- (iii) Also write down the equations of regression lines.

Solution:

$$(i) \quad \text{Here } \bar{X} = 20, \bar{Y} = 45, b_{yx} = 4, b_{xy} = \frac{1}{9}$$

$$\text{Also } r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{4 \times \frac{1}{9}} \Rightarrow r = \frac{2}{3} = 0.667$$

(+ve sign with r is taken because b_{xy} and b_{yx} both are positive).

Hence, $r = 0.667$.

$$(ii) b_{yx} = r \frac{\sigma_y}{\sigma_x} \Rightarrow 4 = \frac{2}{3} \times \frac{12}{\sigma_x} \Rightarrow \sigma_x = 2$$

$$(iii) \text{ Regression equation of } X \text{ on } Y: X - \bar{X} = b_{xy}(Y - \bar{Y})$$

$$\Rightarrow X - 20 = \frac{1}{9} (Y - 45), \text{ or } X = \frac{1}{9} Y + 15$$

Which is the required regression line X on Y .

$$\text{Again regression equation of } Y \text{ on } X: Y - \bar{Y} = b_{xy}(X - \bar{X})$$

$$\Rightarrow Y - 45 = 4(X - 20) \Rightarrow Y = 4X - 35,$$

Which is the required regression equation of Y on X .

Example 2.10. Find the two lines of regression from the following data:

Age of husband		25	22	28	26	35	20	22	40	20	18
Age of wife		18	15	20	17	22	14	16	21	15	14

Hence estimate: (i) the age of husband when the age of wife is 19 and

(ii) The age of wife when the age of husband is 30.

Solution: We have the following table by taking assumed mean A= 26 for X series and assumed mean B= 17 for Y series, Let x= (X-A)= (X-26) and y = (Y-B)= (Y-17).

Table: Computation of Regression coefficients

Age of husband	x	x ²	Age of wife	y	y ²	xy
	X(X-26)			Y (Y-17)		
25	-1	1	18	1	1	-1
22	-4	16	15	-2	4	8
28	2	4	20	3	9	6
26	0	0	17	0	0	0
35	9	81	22	5	25	45
20	-6	36	14	-3	9	18
22	-4	16	16	-1	1	4
40	14	196	21	4	16	56
20	-6	36	15	-2	4	12
18	-8	64	14	-3	9	24
$\sum X$ = 256	$\sum X = 4$	$\sum X^2$ = 450	$\sum Y$ = 172	$\sum Y$ = 172	$\sum Y^2$ = 78	$\sum XY$ = 172

$$\bar{X} - \frac{256}{10} = 25.6$$

$$\bar{Y} - \frac{172}{10} = 17.2$$

Regression coefficient:

$$b_{XY} = r \frac{\sigma_Y}{\sigma_X} = \frac{N \sum xy - \sum x \sum y}{N \sum y^2 - (\sum y)^2}$$

$$= \frac{10 \times 172 - (-4)^2}{10 \times 78 - (+2)^2} = \frac{1720 + 8}{780 - 4} = \frac{1728}{776} = 2.23$$

Regression coefficient:

$$b_{YX} = r \frac{\sigma_X}{\sigma_Y} = \frac{N \sum xy - \sum x \sum y}{N \sum x^2 - (\sum x)^2}$$
$$= \frac{10 \times 172 - (-4)^2}{10 \times 450 - (-4)^2} = \frac{1720 + 8}{4500 - 16} = \frac{1728}{4484} = 0.385$$

The regression line of X on Y is:

$$X - \bar{X} = r \frac{\sigma_x}{\sigma_y} (Y - \bar{Y})$$

$$\Rightarrow X - 25.6 = 2.23(Y - 17.2) = 2.23Y - 38.36$$

$$\Rightarrow X = 2.23Y - 38.36 + 25.6$$

$$\Rightarrow X = 2.23Y - 12.76$$

The regression line of Y on X is $Y - \bar{Y} = r \frac{\sigma_y}{\sigma_x} (X - \bar{X})$

$$\Rightarrow Y - 17.2 = 0.385(X - 25.6) = 0.385X - 9.86$$

$$\Rightarrow Y = 0.385X - 9.86 + 17.2 \quad \text{or} \quad Y = 0.385x + 7.34$$

(i) When $Y=19$ then $X = 2.23 \times 19 - 12.76 = 42.37 - 12.76 = 29.61$ or 30 nearly.

Hence (i) Required estimated age of the husband = 30 years.

(ii) Required estimated age of the wife = 19 years

You may try the following exercises.

12.5 Exercises

E-1 Explain the concepts of correlation and regression. Calculate coefficient of correlation and regression line of Y on X for the data below:

X:	1	2	3	4	5	6
Y:	9	8	10	12	11	13

E-2 Estimate the most probable value of Y when $X = 40$.

In a correlation study, the following values are obtained:

	X	Y
Mean	65	67

Standard Deviation	2.5	3.5
Coefficient of Correlation	0.8	

Find the two regression equation that are associated with the above values.

E-3 Estimate the marks in Mathematics (X) corresponding to 70 marks in English (Y).

Given	X series	Y series
Mean	18	100
Standard Deviation	14	20
Coefficient of Correlation between X and Y series is	+0.8	

E-4 The equation of two lines o regression obtained in correlation analysis are as follows:

$2x+3y-8=0$ and $x+2y-5=0$ and the variance of X =4. Find (i) mean value of x and y, (ii) coefficient of correlation between x and y, (iii) the standard deviation of y, (iv) coefficient of determination of x and y, and (v) the standard error of estimate of regression of x and y.

E-5 From the following data, write down the equation of the regression lines.

	X series	Y series
Arithmetic Mean	60	120
Standard Deviation	3.36	7.24
Coefficient of Correlation	+0.98	

E-6 You are given the following data:

	X series	Y series
Arithmetic Mean	36	85
Standard Deviation	11	8
Coefficient of Correlation between X and Y=		0.66

Obtains the two regression lines

E-7 From the following data of rainfall and production of rice, find the most likely production corresponding to the rainfall 40cms.

	Rainfall	Production (quintals)
Mean	35	50
Standard Deviation	5	8
Coefficient of Correlation	+0.8	

E-8 by using simultaneous equation method, find from the following data

X:	1	2	3	4	5
Y:	2	5	3	8	7

- (a) Regression equation of Y on X;
- (b) (b) Regression equation of X on Y
- (c) The most probable value of Y, when X=10.

12.6 Solutions/ Answers

E-1 $r = 0.886$; $Y = 0.886X + 7.4$

E-2 i) $X = 0.57Y + 26.81$; (ii) $Y = 1.12X - 5.8$

E-3 i) When $Y = 90$ then $X = 12.4$; when $X = 70$ then $Y = 159.28$

E-4(i) $\bar{x} = 1$, $\bar{y} = 2$ (ii) $r = -0.886$ (iii) $\sigma_y = 1.155$ (iv) $r^2 = 0.75$ (v) $S_{xy} = 11$.

E-5 $X = 0.46Y + 4.8$; $Y = 2.11X - 6.6$

E-6 $X = 0.91Y - 41.35$; $X = 0.48X + 67.72$ (ii) When $Y = 75$, then $X = 26.9$

E-7 $Y = 1.2X + 5.2$; when $X = 40$, then $Y = 56.4$ quintals.

E-8 i) $Y = 1.1 + 1.3X$; (ii) $X = 0.5 + 0.5Y$; (iii) $Y = 14.1$.

Assuming the lines of regression of 'y on x' and 'x on y' as: $2y = -2 + 5x$ and $2x = 8 - 3y$

We have $b_{yx} = r \frac{\sigma_y}{\sigma_x} = -\frac{1}{2}$ and $b_{xy} = r \frac{\sigma_x}{\sigma_y} = -\frac{3}{2}$

Multiplying these two regression coefficients we get

$$r^2 = b_{yx}b_{xy} = \left(-\frac{1}{2}\right)\left(-\frac{3}{2}\right) = 0.75 \text{ and } r = \pm 0.866$$

Since b_{yx} and b_{xy} are negative therefore coefficient of correlation r should also be negative.

Hence $R = 0.866$.

If we had assumed the lines regression of y on x on y as

$2x + 3y - 8 = 0$ and $x + 2y - 5 = 0$ then $b_{yx} = -\frac{2}{3}$ and $b_{xy} = -2$

So that $r^2 = b_{yx}b_{xy} = \left(-\frac{2}{3}\right)(-2) = \frac{4}{3} > 1$.

Which is not possible. Our choice is wrong. We should change of choice of lines.

12.7 Summary

The relationship between any two variables may be linear or nonlinear. A relationship may be describing by means of a straight line or a curve. If it is best explained by a straight line. It is called linear regression. If it is described more appropriately by a curve, it is said to be nonlinear regression. There are two regression lines because any one of the two variables may be taken as an independent variable while the other is treated as dependent variable.

12.8 Further Readings

1. Goon, Gupta & Dasgupta : Fundamentals of Statistics Vol. I and Vol. II The World Press Pvt. Ltd., Kolkata.
2. Yule G.U. and Kerdall, M.G.: An Introduction to the Theory of Statistics, Charles Griffin and Company Ltd.

3. C.E. Weatherburn : Mathematical Statistics.

Unit-13 : Rank Correlation and Intra-class Correlation

Structure

13.1 Introduction

13.2 Objectives

13.3 Spearman's Rank Correlation Coefficient

13.4 Rank Correlation Coefficient for tied ranks

13.5 Intra class correlation

13.6 Some remark on Intra class correlation

13.7 Exercises

13.8 Summary

13.9 Further Readings

13.1 Introduction

Karl Pearson's correlation coefficient is a measure for degree of linear relationship between two variables. This correlation coefficient is also known as product moment correlation coefficient. In this calculation of product moment correlation coefficient, It is essential that the two characteristics be definitely measurable. But in many cases the characteristics may not be measurable or even if measurable, may not be measured for lack of measuring instruments. If the characteristic is qualitative such as intelligence or beauty, the scales for the measurements of such traits are not objective and unique. Some times, although measurements may be available for the calculation of product moment correlation coefficient a rough and ready, substitute may still be called for to reduce the mathematical calculations. In all these situations rank correlation coefficient may be used.

Let us suppose that is possible to arrange the n individuals according to the degree to which they possess the characteristics under inquiry although the characteristic may not be directly measurable. Thus, for example a number of operators it may not be easy to offer some numerical measure of efficiency. Such an ordered measurement be called a ranking and the ordinal number indicating the position of a given individual in the ranking is called its rank. For example, the marks obtained by five students are 53, 89, 17, 67 and 35; their respective ranks are 3, 1, 5, 2 and 4. A ranking in which two or more occupying the same position (Rank) is merit or proficiency is possession of two characteristic A and B. The ranks in the two characteristics will in general be different. For example, if we consider the relation between intelligence and beauty, it is not necessary that a beautiful individual is intelligent also. Let x_i and y_i be the ranks of i^{th} individual in two characteristics A and B respectively. The product moment correlation coefficient between the ranks x_i and y_i 's is called the rank correlation coefficient. The formula for rank correlation coefficient is derived by Spearman. Thus, it is called Spearman's rank correlation coefficient.

13.2 Objectives

After going through this unit, you should be able to

- Define Rank Correlation Coefficient
- Describe intra class correlation coefficient.

13.3 Spearman's Rank Correlation Coefficient

First, let us suppose that there is no tie, i.e., no two individuals are ranked equal in either variable. The ranks x 's and y 's take values $1, 2, 3, \dots, n$ in some order. Hence

$$\sum x_i = \sum y_i = 1 + 2 + \dots + n(n+1)/2.$$

and means are $\bar{x} = \bar{y} = \frac{n+1}{2} \dots \dots \dots (3.1)$

Sum of squares are given as

$$\sum x_i^2 = \sum y_i^2 = 1^2 + 2^2 + \dots + n^2(n+1)(2n+1)/6.$$

and $\sigma_x^2 = V(x) = \left(\frac{1}{n} \sum X^2 - \bar{X}^2 \right)$

$$= \left\{ \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2} \right)^2 \right\}$$
$$= \frac{n^2 - 1}{12} - V(Y) = \sigma^2 \dots \dots \dots (3.2)$$

Let $d_i = x_i - y_i$,

Since $\bar{x} = \bar{y}$, d_i can be written as $d_i = (x_i - \bar{x}) - (y_i - \bar{y})$

Squaring both sides and summing over i from 1 to n we get

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

Dividing both sides by n sides n we have

$$\frac{1}{n} \sum_{i=1}^n d_i^2 = \sigma_x^2 + \sigma_y^2 - 2Cov(x, y) \dots \dots \dots (3.3)$$

Let ρ be the correlation coefficient between the ranks x and y, then

$$\rho = \frac{Cov(x, y)}{\sigma_x \sigma_y} \text{ or } Cov(x, y) = \rho \sigma_x \sigma_y \dots \dots \dots (3.4)$$

From (3.3) and (3.4)

$$\frac{1}{n} \sum_{i=1}^n d_i^2 = \sigma_x^2 + \sigma_y^2 - 2\rho \sigma_x \sigma_y$$

Since $\sigma_x^2 = \sigma_y^2$, we have

$$\frac{1}{n} \sum_{i=1}^n d_i^2 = 2\sigma_x^2 - 2\rho \sigma_x^2$$

$$= 2\sigma_x^2 (1 - \rho)$$

$$(1 - \rho) = \frac{\sum_{i=1}^n d_i^2}{2n\sigma_x^2}$$

or

$$\rho = 1 - \frac{\sum d_i^2}{2n\sigma_x^2}$$

Putting the value of σ_x^2 from (3.2) we get

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \dots \dots \dots (3.5)$$

Which is Spearman's formula for the rank correlation coefficient. Since the rank correlation coefficient is simple product moment correlation coefficient between tow series of ranks, it always lies between -1 to +1.

For perfect agreement $x_i = y_i$ for each i and $d_i = 0$ for all i,

$$\text{So } \sum d_i^2 = 0 \text{ and } \rho = 1.$$

For perfect disagreement $y_i = n - x_i + 1 \quad \forall i$

$$d_i = x_i - y_i = x_i - (n - x_i + 1)$$

$$= 2x_i - (n + 1)$$

$$\sum d_i^2 = 4 \sum_{i=1}^n \left(x_i - \frac{n+1}{2} \right)^2 = 4n\sigma_x^2 = \frac{n(n^2 - 1)}{3}$$

$$\rho = 1 - \frac{\frac{n(n^2-1)}{3}}{n(n^2-1)} = 1 - 2 = -1.$$

13.4 Rank Correlation Coefficient for Tied Ranks

If some of the individuals have the same rank in ranking they are said to be tied. If the same rank is allocated to m individuals, then we have a tie of length m . Let us suppose that m individuals, say, $(k+1)^{\text{th}}, (k+2)^{\text{th}}, \dots, (k+m)^{\text{th}}$ are tied. If these individuals follow k other individuals in the ranking then each may be given the rank $(k+1)$.

But we shall follow the convention that each of the k individuals is to be given the rank

$$\frac{(k+1)(k+2) + \dots + (k+n)}{m} = k + \frac{m+1}{2} \dots \dots \dots (3.6)$$

This is the arithmetic mean of the ranks that these individuals would have received had there been no ties.

This then does not affect the mean of the ranks. However this will affect the variance. The sum of squares of untied ranks would be.

$$(k+1)^2 + (k+2)^2 + \dots + (k+m)^2 \\ Mk^2 + m(m+1)k + \frac{1}{6}m(m+1)(2m+1) \dots \dots \dots (3.7)$$

and the sum of square of the tied ranks is

$$m \left\{ k + \frac{m+1}{2} \right\}^2 = mk^2 + m(m+1)k + \frac{1}{4}m(m+1)^2 \dots \dots \dots (3.8)$$

The difference being $(m^3-m)/12$. Consequently, the variance is lowered by $(m^3-m)/12n$ in the case of tied rank. It is obvious that the effect of different tied sets are additive. Suppose that there are s such sets of ranks to be tied in the x -series of length $m_1, m_2, \dots, m_s$ and in the ranking with respect to the second characteristic Y , there are t ties of length $m_1, m_2, \dots, m_t$

The variance would then be

$$\sigma_x^2 = \frac{n-1}{12} - T_x$$

and

$$\sigma_y^2 = \frac{n-1}{12} - T_y$$

Where

$$T_x = \sum_{i=1}^n \frac{(m_i^3 - m_i)}{12n}$$

and

$$T_y = \sum_{j=1}^n \frac{(m_j^3 - m_j)}{12n}$$

Similarly, from (3.3), $2Cov(x, y) = \sigma_x^2 + \sigma_y^2 - \frac{\sum d_i^2}{2n}$

The variance for the case of tied ranks would be

$$Cov(x, y) = \frac{n^2-1}{12} - \frac{T_x + T_y}{2} - \frac{\sum d_i^2}{2n}$$

So the Spearman's rank correlation coefficient in the case of tied ranks becomes

$$\rho = \frac{\frac{n^2-1}{12} - \frac{T_x+T_y}{2} - \frac{1}{2n}\sum d_i^2}{\left(\frac{n^2-1}{12} - T_x\right)\left(\frac{n^2-1}{12} - T_y\right)} \dots\dots\dots(3.9)$$

In the case of perfect agreement between the two sets of ranks $x_i = y_i$ for all i , then

$T_x = T_y$ and hence

$$\rho = \frac{\left(\frac{n^2-1}{12} - T_x\right)}{\left(\frac{n^2-1}{12} - T_x\right)} = 1$$

In the case of perfect disagreement we have $y_i = n-x_i + 1$ for each i , therefore,

$$\sigma_y = \sigma_x = \frac{n^2-1}{12} - T_x$$

Since $T_x = T_y$ in this case also.

Further
$$\frac{1}{n} \sum_{i=1}^n d_i^2 = \sum_i [2x_i - (n+1)]^2 / n$$
$$= 4\sigma_x^2 = \frac{n^2 - 1}{3} - 4T_x$$

And
$$\rho = \frac{\frac{n^2-1}{12} - T_x - \left(\frac{n^2-1}{6} - 2T_x\right)}{\frac{n^2-1}{12} - T_x}$$

Remarks on Spearman's rank correlation:

(i) $i \sum d_i = \sum (x_i - y_i) = \sum x_i - \sum y_i = 0$

This provides a check for numerical calculations.

(ii) Since rank correlation coefficient ρ is nothing but product moment correlation coefficient between the ranks it can be interpreted in the same way as product moment correlation coefficient.

(iii) Karl Pearson's coefficient is meaningful if observation are linearly (approximately) related otherwise it may be meaningless.

(iv) Spearman's rank correlation coefficient is distribution free measure.

(v) Spearman's coefficient is used for finding correlation coefficient if we are dealing with qualitative characteristic which cannot be measured quantitatively but can be arranged in order of magnitude. It can also be used where both characteristics are quantitative.

(vi) If we adjust only the covariance term i.e. $\sum x_i y_i$, and not the variances σ_x^2 (or $\sum x_i^2$) and σ_y^2 (or $\sum y_i^2$) for ties then the formula (3.9) reduces to

$$\rho = 1 - \frac{6 \left[\frac{1}{n} \sum d_i^2 + T_x + T_y \right]}{(n^2 - 1)} \dots \dots \dots (3.10)$$

This is generally used in practice for computation of rank correlation coefficient for simplification.

Check your progress

1. Define the rank correlations.

2. What do you mean by Spearman's rank correlation?

Example:1: Ten hand writing were ranked by two judges in a completion. The ranking are given below. Calculate Spearman's rank correlation coefficient.

Hand writing										
	A	B	C	D	E	F	G	H	I	J
Judge-I	3	8	5	4	7	10	1	2	6	9
Judge-II	6	4	7	5	10	3	2	1	9	8

Solution: The differences of ranks between the ranks $d_i = x_i - y_i$ of two judges for ten observation are -3, 4, -2, -1, -3, 7, -1, 1, -3, 1

Hence $\sum d_i^2 = 9 + 16 + 4 + 1 + 9 + 49 + 1 + 1 + 9 + 1 = 100$

Thus Spearman's rank correlation coefficient

$$= 1 - \frac{6 \times 100}{10(10^2 - 1)} = 1 - \frac{600}{900} = 0.394 \text{ answer}$$

Example:2: Ten competitors in a musical test were ranked by the three judges A,B and C in the following order

Rank by A: 1 6 5 10 3 2 4 9 7 8

Rank by B: 3 5 8 4 7 10 2 1 6 9

Rank by C: 6 4 9 8 1 2 3 10 5 7

Using rank correlation coefficient discuss which pair of judges has the nearest approach common likings in music.

Solution:

Rank by A	Rank by B	Rank by C	$d_1^2 = (X_i - y_i)^2$	$d_2^2 = (X_i - z_i)^2$	$d_3^2 = (X_i - z_i)^2$
1	3	6	4	25	9
6	5	4	1	4	1
5	8	9	9	16	1
10	4	8	36	4	16
3	7	1	16	4	36
2	10	2	64	0	64
4	2	3	4	1	1
9	1	10	64	1	81
7	6	5	1	4	1

8	9	7	1	1	4
			$\sum d_i^2 = 200$	$\sum d_2^2 = 60$	$\sum d_3^2 = 214$

$$\rho(x, y) = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \times 200}{10(100 - 1)} = -\frac{7}{33}$$

$$\rho(x, z) = 1 - \frac{6 \times 600}{10 \times 99} = \frac{7}{11}$$

$$\rho(y, z) = 1 - \frac{6 \times 214}{10 \times 99} = -\frac{49}{165}$$

Since $\rho(x, z)$ is maximum, we conclude the judges A and C have the nearest approach to common likings in music.

Example: 3: Two supervisors ranked 12 workers working under them in order of efficiency as follows.

No. of workers	1	2	3	4	5	6	7	8	9	10	11	12
Supervisor-I	5	6	1	2	3	8.5	8.5	4	7	11	10	12
Supervisor-II	5.5	5.5	2	2	2	9	7	4	8	10.5	12	10.5

Calculate rank correlation coefficient.

Solution: In the ranking of first supervisor, there 15 one tie of length 2.

Thus
$$T_x = \frac{1}{12} \frac{2^3 - 2}{12} = 0.0417$$

In the ranking of second supervisor, there are three ties of length 2, 3 and 2 respectively. Thus

$$T_y = \frac{1}{12} \left[\frac{2^3 - 2}{12} + \frac{3^3 - 3}{12} + \frac{2^3 - 2}{12} \right] = 0.25$$

$$\sum d_i^2 = .25 + .25 + 1 + 0 + 1 + .25 + 2.25 + 0 + 1 + .25 + 4 + 2.25 = 12.50$$

$$\rho = \frac{\frac{n^2 - 1}{12} - \frac{T_x + T_y}{2} - \frac{1}{2n} \sum d_i^2}{\left(\frac{n^2 - 1}{12} - T_x \right)^{1/2} \left(\frac{n^2 - 1}{12} - T_y \right)^{1/2}}$$

$$= \frac{\frac{12^2 - 1}{12} - \frac{0.0417 + 0.25}{2} - \frac{12.50}{2 \times 12}}{\sqrt{\frac{12^2 - 1}{12} - 0.0417} \sqrt{\frac{12^2 - 1}{12} - 0.25}}$$

$$= \frac{11.25}{\sqrt{11.8750} \times \sqrt{11.6667}} = 0.956 \text{ Answer.}$$

Example 4: A sample of 10 fathers and their eldest sons gave the following data about their height in inches.

Father: 65, 63, 67.5, 64, 68, 62, 70, 66, 66, 68.5, 71

Son: 68, 66, 68.5, 65.5, 69, 66.5, 69.5, 65, 71, 70

Calculate Spearman's rank correlation coefficient.

67.5	68.5	5	5	0	0
64	65.5	8	9	-1	1
68	69	4	4	0	0
62	66.5	10	7	3	9
70	64.5	2	3	-1	1
66	65	6	10	-4	16
68.5	71	3	1	2	4
71	70	1	2	-1	1
				$\sum d_i = 0$	$\sum d_i^2 = 34$

Spearman's rank correlation is given by

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

$$= 1 - \frac{6 \times 34}{10 \times 99} = 1 - .2060 = 0.7940 \text{ andwer}$$

13.5 Intra-Class Correlation

The product moment correlation coefficient is defined between two clearly defined variables such as height and weight. There are cases in which we require the correlation with respect to a particular variate between members of the same class. For example between heights of brothers in the same family, intra class correlation means within class correlation. It is distinguishable from

product moment correlation coefficient in as much as here both the variables measure the same characteristic. Sometimes specially in biological and agricultural study, it is of interest to know how the members of a family or group are correlated among themselves with respect to some one of their common characteristics. For example we may require the correlation between the heights of brothers of a family. In such cases both variables measures the same characteristics i.e., height and height. There is nothing to distinguish one from the other so that one may be treated as x-variables and the other may as y variable correlation. It is distinguishable from product moment correlation coefficient in as much as here both the variables measure the same characteristic. Sometimes specially in biological and agricultural study, it is of interest to know how the members of a family or group are correlated among themselves with respect to some one of their common characteristics. For example, we may require the correlation between the heights of brothers of a family. In such cases both variables measure the same characteristic i.e., height and height. There is nothing to distinguish one from the other so that one may be treated as x-variable and the other may as y variable.

Suppose we have n classes with $d_1, d_2, \dots, d_n$ members each of which have values as

$$\begin{array}{c} X_{11}, X_{12}, \dots, X_{n1} \\ X_{12}, X_{22}, \dots, X_{n2} \\ \vdots X_{22}, \dots, \vdots \\ \vdots X_{22}, \dots, \vdots \\ \vdots X_{22}, \dots, \vdots \\ X_{1k1}, X_{2k2}, \dots, X_{nkn} \end{array}$$

Here x_{ij} ($i=1,2,\dots,n$; $j=1, 2, 3, \dots, k_i$) denote the measurement of j^{th} member of i^{th} class.

We shall have $k_i(k_i - 1)$ pairs for i^{th} class or group like (X_{ij}, X_{il}) $j \neq l$. There will be $\sum k_i(k_i - 1)$ pairs in total for all n classes. If we prepare a correlation table there will be $k_i(k_i - 1)$ entires for the i^{th} class and $N = \sum k_i(k_i - 1)$ entires for all the n classes, such a table is called an intra class correlation table and correlation is called the intra class correlation i.e., intra-class correlation is product moment correlation coefficient between all $N = \sum k_i(k_i - 1)$ pairs of observations.

In the intra class correlation table (bivariate table) X_{i1} occurs (k_i-1) times, x_{i2} occurs (k_i-1) times and so on from the i^{th} class. Hence the total for i^{th} is $(k_i-1) \sum_j X_{ij}$ and total for all n classes is $\sum (k_i - 1) \sum_j X_{ij}$

Thus means are $\bar{x} = \bar{y} = \frac{1}{N} [\sum (k_i - 1) \sum_j X_{ij}]$

Similarly $\sigma_x^2 = \sigma_y^2 = \frac{1}{N} \left[\sum (k_i - 1) \sum_j (x_{ij} - \bar{x})^2 \right]$

and $Cov(x, y) = \frac{1}{N} \sum_{i=1} \left[\sum_{j \neq 1} \sum (x_{ij} - \bar{x})(x_{il} - \bar{x}) \right]$

$$= \frac{1}{N} \sum_{i=1} \left[\sum_{j \neq 1} \sum (x_{ij} - \bar{x})(x_{il} - \bar{x}) - \sum_j (x_{ij} - \bar{x})^2 \right]^{(AB) = \frac{(A)(B)}{N}}$$

If we write $\bar{x}_i = \frac{1}{k} \sum_j x_{ij}$ then

$$\begin{aligned} &= \sum_{i=1} \left[\sum_j \sum (x_{ij} - \bar{x})(x_{il} - \bar{x}) \right] \\ &= \sum_i k_i (\bar{x}_i - \bar{x}) k_i (\bar{x}_i - \bar{x}) \\ &= \sum_i k_i^2 (\bar{x}_i - \bar{x})^2 \end{aligned}$$

Therefore intra class correlation coefficient is given by

$$\begin{aligned} r &= r(X, y) = \frac{Cov(x, y)}{\sigma_x \sigma_y} \\ &= \frac{\sum_i k_i^2 (\bar{x}_i - \bar{x})^2 - \sum_i \sum_j (x_{ij} - \bar{x})^2}{\sum_i \sum_j (k_i - 1) (x_{ij} - \bar{x})^2} \dots \dots (3.11) \end{aligned}$$

If we put $k_i = k$ i.e. all the n classes has equal members k then

Range of intra class correlation can be obtained from (3.12)

$$\begin{aligned} 1 + (k - 1)r &= \frac{k\sigma_m^2}{\sigma^2} \geq 0 \\ \Rightarrow (k-1)r &\geq -1 \\ \Rightarrow r &\geq -\frac{1}{k-1} \end{aligned}$$

Since r is correlation coefficient it cannot be greater than 1 thus range of intra class correlation is

$$-\frac{1}{k-1} \leq r \leq 1.$$

13.6 Some Remarks on Intra-Class Correlation

Note that $0 \leq \sigma_m^2 \leq \sigma^2$. Hence this coefficient is a maximum i.e., equal to 1 if $\sigma_m^2 = \sigma^2$ i.e., when variance between the means is equal to the total variance which will happen when the variance

within class is zero. In this case the variate values for members within each class are all equal. Again r is minimum i.e. equal to $-\frac{1}{k-1}$ if $\sigma_m^2=0$ which happens when the variance within classes is the maximum possible. Thus the coefficient of intra-class correlation may be looked upon as a measure of the extent to which the total variance is explained away by the variance between the means.

Since intra class correlation cannot be less than $-\frac{1}{k-1}$ through it may attain the value $+1$ on the positive side, so *it is a skew coefficient and a negative value has not the same significance as a departure from independence as an equivalent positive value*. Intra class correlation is unaffected by change of origin and scale.

Again r is minimum i.e., equal to $-\frac{1}{k-1}$ if $\sigma_m^2 = 0$ which happens when the variance within classes is the maximum possible. Thus the coefficient of intra class correlation may be looked upon as a measure of the extent to which the total variance is explained away by the variance between the means.

Since intra-class correlation cannot be less than $-\frac{1}{k-1}$, thorough it may attain the value $+1$ on the positive side, so *it is a skew coefficient and a negative value has not the same significance as a departure form independence as an equivalent positive value*. Intra class correlation is unaffected by change of origin and scale.

Example: The weights in gm of a number of copper wires each of length one meter, were obtained. These are shown below classified according to the die from which they come. Determine intra class correlation.

Die No.				
I	II	III	IV	V
1.33	1.30	1.32	1.31	1.30
1.32	1.35	1.29	1.29	1.32
1.36	1.33	1.31	1.33	1.33
1.35	1.35	1.28	1.31	1.33

Solution:

Since intra class correlation coefficient is invariant under change of origin and scale, let $u_{ij} = 100(X_{ij} - 1.28)$ the change values are given in following table

Die No.				
I	II	III	IV	V
5	2	4	3	2
4	7	1	1	4
8	5	3	5	5
7	6	0	3	5
Total: 24	20	8	12	16
Mean: 6	5	2	3	4

$$\bar{u} = 80/20 = 4$$

$$\begin{aligned}\sigma^2 &= \frac{1}{20} \sum_i \sum_j (u_{ij} - \bar{u})^2 \\ &= \frac{1}{20} [1 + 0 + 16 + 9 + 4 + 9 + 1 + 4 + 0 + 9 + 1 + 16 + 1 + 9 + 1 + 4 + 0 + 1 + 1] \\ &= \frac{88}{20} = 4.4\end{aligned}$$

$$\sigma_m^2 = \frac{1}{5} [4 + 1 + 4 + 1 + 0] = 2$$

$$r = \frac{1}{k-1} \left[\frac{k\sigma_m^2}{\sigma^2} \right] = \frac{1}{3} \left[\frac{4 \times 2}{\sigma^2 4.4} - 1 \right] = 0.273 \text{ answer.}$$

13.7 Exercises

E-1 (a) Define rank correlation coefficient Explain the difference between product correlation coefficient and rank correlation coefficient.

(b) Prove that the spearman's rank correlation coefficient is given by where d_i denotes the difference between the ranks of i^{th} individual.

E-2 The rankings of ten students in two subjects A and B are as follows:

A: 3 5 8 4 7 10 2 1 6 9
 B: 6 4 9 8 1 2 3 10 5 7

Find the correlation coefficient.

E-3 Ten recruits were subjected to a selection test to ascertain their suitability for a certain course of training. At the end of training they were given a proficiency test.

The marks secured by recruits in the selection test (X) and the proficiency test (Y) are given below:-

Serial No. :	1	2	3	4	5	6	7	8	9	10
A:	10	15	12	17	13	16	24	14	22	20
B:	30	42	45	46	33	34	40	35	39	38

Calculate product moment correlation coefficient and rank correlation coefficient. Why are two coefficient different?

E-4 Obtain the formula for rank correlation coefficient in case of ties.

E-5 The I.Q.'s of group of six persons were measured and they then sat for a certain examination. Their I.Q.'s and examination marks were as follows:

Person :	A	B	C	D	E	F	
I.Q.	:	110	100	140	120	80	90
Exam marks:	70	60	80	60	10	20	

Compute the coefficient of correlation and rank correlation. Why are the correlation figures obtained different?

E-6 Ten competitors in a beauty contest ranked by three judges as follows:

Competitors										
Judges:	1	2	3	4	5	6	7	8	9	10
A:	6	5	3	10	2	4	9	7	8	1
B:	5	8	4	7	10	2	1	6	9	3
C:	4	9	8	1	2	3	10	5	7	6

Discuss which pair of judges has the nearest approach to common tastes of beauty.

E-7 A sample of 12 fathers and their eldest sons gave the following data about their heights in inches:

Father:	65	63	67	64	68	62	70	66	68	67	69	71
Son:	68	66	68	65	69	66	68	65	71	67	68	70

Calculate coefficient of rank correlation.

E-8 Define intra-class correlation. Derive the formula for the intra class correlation when the variable x is observed for p families, each consisting of k members.

E-9 Show that the coefficient derived in previous question lies between the limits $-\frac{1}{k-1}$ and +1.

E-10. For each of six families, the heights in inches of three brothers belonging to it are recorded below. Compute the coefficient of intra class correlation.

Family	Heights of brothers		
1	69.5	70.6	72.3
2	71.2	70.8	72.0
3	65.6	67.2	66.7
4	62.2	63.6	63.5
5	68.0	70.5	70.5
6	64.5	64.3	64.6

13.8 Summary

While dealing with qualitative characteristics like intelligence, it is advisable not to use the actual measurements for the calculation of correlation coefficient. Not only that different examiners in such a case will award different marks for the same intelligence, the same examiner at two different times may award different marks for the same intelligence. In such cases, therefore ordinal numbers are allotted to actual measurements either in ascending or descending order. These ordinal numbers are called ranks. The correlation coefficient between the ranks of two such variables or characteristics is known as rank correlation coefficient.

The correlation coefficient between the members of same class is known as intra class correlation coefficient.

13.9 Further Readings

1. Goon , Gupta & Dasgupta : Fundamentals of Statistics Vol. I and Vol. II The World Press Pvt. Ltd., Kolkata.
2. Yule G.U. and Kerdall, M.G.: An Introduction to the Theory of Statistics, Charles Griffin and Company Ltd.
3. C.E. Weatherburn : Mathematical Statistics.

Unit-14 : Theory of Attributes

Structure

- 14.1 Introduction**
- 14.2 Objectives**
- 14.3 Combinations, Classes and Class frequencies of Attributes**
- 14.4 Consistency of Data**
- 14.5 Joint Distribution of Attributes (i.e. Contingency tables)**
- 14.6 Independence and Association of Attributes**
- 14.7 Measures of Association**
- 14.8 Yates Correction**
- 14.9 Exercises**
- 14.10 Summary**
- 14.11 Further Readings**

14.1 Introduction

In real life we often come across situation where actual measurements (in numerical terms) of characteristics under study are not possible. Simply we can observe the presence or absence of these specific characteristics (or properties) over a group of individuals or units under observation. Such characteristics are termed as attributes. Thus attribute is a qualitative character which can not be measured numerically and one simply observes the presence or absence of it. Honesty, beauty, preferences, likings, colors, blindness, smoking etc are few examples of attributes. If an attribute is classified in two groups it is called a dichotomous attribute whereas if it is classified in many categories it is called manifold.

14.2 Objectives

After going through this unit, you should be able to

- Define attributes
- Enumerate classification of attributes
- Define consistency of data
- Define independence and association of attributes
- Discuss measures of association for 2×2 case
- City Yates correction

14.3 Combinations, Classes and Class Frequencies of Attributes (Dichotomous Classification)

Different attributes, their subgroups and combinations are called classes and the number of observation assigned to them are called respective class frequencies. Conventionally A,B, C,...etc. denote positive attribute or presence of attribute $\alpha, \beta, \gamma \dots etc$ and etc denote the negative attribute or absence of attribute. Thus α is equivalent to the class none of which possess attribute A. It is to be noted that object must belong to one and only one subgroup of an attribute. For example, either to A or to α .

Example.1 Consider a dichotomous attribute “Smoking” which may be divided into two categories

1. Smokers (A)
2. Nonsmokers (α)

Let us further consider one more dichotomous attribute “Literacy” which is also classification into two categories

1. Literate (B)
2. Illiterate (β)

Both of these attributes are called dichotomous as they are divided into two subgroups and any individual will belong to only one of these categories either A or α and B or β .

Symbolically we may write

(AB) to denotes the number of individuals who are smokers and literate.

(A β) to denote the number of individuals who are smokers and illiterate.

(α , B) to denote the number of individuals who are nonsmokers and illiterate.

(α, β)to denote the number of individuals who are non smokers and illiterate.

Thus in general (AB) stands for the frequency of the individuals or units which possess the attribute A and B simultaneously. Similar interpretations exists (α, β) for etc. Total number of observation is denoted by N.

Obviously we have,

$$N = (A) + (\alpha) = (B) + (\beta)$$

$$N = (AB) + (A\beta) + (\alpha B) + (\alpha\beta)$$

Total number of observation N is taken as one class of order “Zero”. (A), (B), (C),..... etc will be each of order one, $(AB) + (A\beta) + (A\beta) + (\alpha\beta) \dots$ etc. will be each of order three and so on. A class or class frequency (ABC...M, r letters) containing r attributes in it said to be of order “r”.

Total number of Class Frequencies with n number of Attributes

Theorem .1: With n number of attributes defined over a group of individuals or units we have 3^n total of classes or class frequencies.

Proof: Let us assume that there be n attributes A,B, C,,M.

1st order frequency: ${}^nC_{1.2}$, because for each of the nC_1 first order positive classes there would be two class frequencies. For example, A will be divided into A and α , B will be divided into B and β and so on....

2nd order frequency: ${}^nC_2.2^2$ because for each of the nC_2 combinations there would be $2^2=4$ classes. For example with A and B would have AB, $A\beta$, $A\beta$, $\alpha\beta$ and so on.....

.

.

.

In general rth order frequency will be ${}^nC_n. 2^r$ in numbers.

and nth order frequency will be ${}^nC_n. 2^n$ in numbers.

therefore the total number of class frequencies with n attributes

$$=1+ {}^nC_{1.2} + {}^nC_{2.2^2} + \dots \dots \dots {}^nC_n.2^n$$

$$= (1 + 2)^n = 3^n$$

Implying that with n number of attributes we have 3^n total number of classes or class frequencies.

Example.2 Let us consider 3 attributes then we have $3^3=27$ classes and corresponding 27 class frequencies.

Order Zero: N, total number of class is 1

Order One: (A), (α), (B), (β), C, γ , total number of classes in this case is 6

Order Two: (AB), (AC), (BC), ($A\beta$), ($\alpha\beta$), ($\alpha\gamma$), ($\beta\gamma$), (αC), (βC), ($A\gamma$), ($B\gamma$), (αB)

Total number of classes in this case is 12

Order three: (ABC), ($A\beta C$), ($AB\gamma$), (αBC), ($\alpha\beta C$), ($\alpha\beta\gamma$), ($\alpha B\gamma$), ($A\beta\gamma$),

Total number of classes in this case is 8

Number or frequencies= frequencies of (order 1+ order 2+ order 3)

$$=1+6+12+8= 27$$

Positive and Negative Attributes and their class frequencies

Attribute denoted by capital letters ABCD.,are termed as positive attributes and those denoted by Greek letters $\alpha\beta\gamma$... are termed as negative attributes. A,AB, ABC..... are positive classes whereas a, $\alpha\beta$, $\alpha\beta\gamma$,.....are negative classes.

Ultimate Class frequencies

Class frequencies of highest order are called ultimate class frequencies. In general with n attributes n being the highest order of the class, class frequencies of order will be called ultimate class frequencies. In particular with three attributes A, B and C the class frequency of order three like (ABC), (AB γ), (α BC) are ultimate class frequencies.

Symbols and Formulae

A symbol AN is taken for dichotomizing N according to attribute A and is written AN= (A).

Therefore we have symbolically,

$$(A)= A.N (\neq \text{product of A and N})$$

= number of objects possessing the attribute A

$$\text{Similarly } (\alpha) = \alpha.N (\neq \text{product of } \alpha \text{ and N})$$

= number of objects not possessing the attribute A

Adding these two expressions, we get

$$(A)+ (\alpha)= AN+ \alpha N$$

$$N= (A+\alpha) N$$

$$\Rightarrow A+\alpha \equiv 1 \text{ or } \alpha= 1-A$$

Thus the operator A or α can be replaced by 1- α or 1-A respectively.

$$\text{Similarly we may take } AB.N= (AB) \text{ or } \alpha \beta N = (\alpha\beta)$$

$$\text{Thus } (\alpha\beta)= \alpha\beta.N = (1\sim A)(1 - B)N$$

$$= (1-A-B+AB)N$$

$$= N-(A)-(B)+(AB)$$

$$\text{Again } (\alpha\beta\gamma)= 1-A) (1-C) N= (1-A-B-C+AB+BC+AC-ABC) N$$

$$= N- (A)-(B)-(C)+(AB)+(BC)+(AC)-(ABC)$$

Example.3: Given that (A) =(α)= (B)= (β)=(C)= (γ)= N/2

$$\text{And } (ABC)= (\alpha\beta\gamma)$$

$$\text{Show that } 2 (ABC)= (AB)+(AC)+(BC)- N/2$$

$$\text{Solution: Since } (\alpha\beta\gamma)= (1-A) (1-B) (1-C)N$$

$$= N - (A) - (B) - (C) + (AB) + (BC) + (AC) - (ABC)$$

$$\therefore (\alpha\beta\gamma) + (ABC) = N - (A) - (B) - (C) + (AB) + (BC) + (AC)$$

$$\text{Further } (\alpha\beta\gamma) = (ABC)$$

$$\text{Therefore, } 2(ABC) = N - 3/2 N + (AB) + (BC) + (AC)$$

$$\text{Or } 2(ABC) = (AB) + (BC) + (AC) - N/2. \quad \text{Proved}$$

Theorem.2: With n attributes there are in all 2^n positive class frequencies.

Proof: Suppose there are A,B,C,...M n attributes is the only class frequency of zero order.

Frequency of first order class = nC_1

Frequency of second order class ${}^nC_{2.1}$ (a combination of any two attributes A and B will give only one positive class frequency (AB),

Frequency of third order class ${}^nC_{3.1}$ (a combination of any two attributes A B,C will give only one positive class frequency (ABC)

.

.

Similarly

Frequency of r^{th} order class = ${}^nC_{r.1}$

.

.

Frequency of n^{th} order class = ${}^nC_{n.1}$

So the total no of class frequencies with n attributes

$$= 1 + {}^nC_1 + {}^nC_2 + \dots + {}^nC_r + \dots + {}^nC_n$$

$$= (1+1)^n = 2^n \quad \text{Proved}$$

Corollary: With n attributes the number of negative class frequencies will be $(2^n - 1)$, because except N. i.e. Zero order class frequency the counter part of every positive class frequency is there in the set of negative class frequencies such as α in place of AB, $\alpha\beta\gamma$ in place of ABC and so on.

Theorem .3: With n attributes ultimate class frequencies will be $2n$ in numbers.

Proof: With n attributes ultimate class frequencies will be of order n and each place can be occupied in two ways in (ABC...M). For ex 2^{nd} place may be B or β , 3^{rd} place may be C or γ and so on. Thus n places with respect to n attributes will create $(2.2.2 \dots 2n \text{ times}) = 2^n$ ultimate class frequencies.

Relations between Class frequencies of different orders:

Every class frequency can be expressed in terms of ultimate class frequencies so the set of 2^n ultimate class frequencies are sufficient for obtaining all the 3^n class frequencies. However none of these 2^n positive or ultimate class frequencies is expressible in terms of rest of the frequencies of that set. In that sense they remain mutually independent. Such a set 2^n independent set is able to specify whole of the data.

14.4 Consistency of Data

A data is said to be consistent if all of its class frequencies have been appeared to have been observed within one and the same population. Class frequencies of a consistent data will not have any mutual contradiction rather they will support each other.

Condition for the consistency:

The necessary and sufficient condition for the consistency of data is that none of the ultimate class frequencies should be negative. The minimum possible value of each of the ultimate classes frequencies is zero.

Example.4: 50% of items have characteristic A and B, 35% have A but not B, 25 % have B but not A-show that there must have been one misprint in the report.

Solution: Let us take $N = 100$, then

In symbols we are given that $(AB) = 50$, $(A\beta) = 35$, $(\alpha B) = 25$

$$(A) = (AB) + (A\beta) = 50 + 35 = 85$$

$$(B) = (AB) + (\alpha B) = 50 + 25 = 75$$

$$\begin{aligned} (\alpha\beta) &= (1-A)(1-B)N = (1-A-B+AB)N \\ &= N - (A) - (B) + (AB) = 100 - 85 - 75 + 50 = -10 \end{aligned}$$

The negative value of the ultimate class frequency shows that the data are inconsistent and therefore there must have been some misprint in the reporting.

Mathematical Conditions for Consistency

With two attributes A and B ultimate class frequencies are (AB) , (A, β) , $(A\beta)$, $(\alpha\beta)$. For the consistency of data, $(AB) \geq 0$(i)

$$(A\beta) \geq 0 \Rightarrow A(1 - B)N \geq 0 \Rightarrow (A - AB)N \geq 0 \Rightarrow (A) - (AB) \geq 0$$

$$\text{or } (AB) \leq (A) \dots \dots \dots (ii)$$

$$\text{Similarly } (\alpha\beta) \geq 0 \Rightarrow (AB) \geq (A) + (B) - N \dots \dots \dots (iii)$$

$$(\alpha\beta) \geq 0 \Rightarrow (B) \Rightarrow (AB) \leq (B) \dots \dots \dots (iv)$$

Example.5: If, in a series of houses actually invaded by small pox 70% of the inhabitants are attacked and 85% have been vaccinated. What is the lowest % of the vaccinated that have been attacked?

	(B) Vaccinated	(β)Not vaccinated	
(A) Attacked	AB	($A\beta$)	70
(B) Not attacked	(βB)	($\alpha\beta$)	30
	85	15	100

Solution: Given that,

$$(A)=70 \quad (B) = 85 \quad (\alpha)= 30, \quad (\beta)= 15, \quad N= 100$$

$$\text{Since, } (AB) \geq 0$$

$$\text{and } (\alpha\beta) \geq 0 \Rightarrow (1-A)(1-B)N \geq 0$$

$$\text{or } (1-A-B+AB)N \geq 0$$

$$\text{or } N - (A) - (B) + (AB) \geq 0$$

$$\text{or } (AB) \geq (A) + (B) - N$$

$$\text{or } (AB) \geq (70) + (85) - 100$$

$$\text{or } (AB) \geq 55$$

So, out of 85 vaccinated at least 55 are being attacked.

$$\text{Hence at least } \frac{55}{85} \times 100 = 64.7\% \text{ vaccinated have been attacked.}$$

Solution: Mathematically,

$$\text{If } (A) = (AB) \text{ and } (B) = (BC)$$

$$\text{Then to prove that } (A) = (AC)$$

$$\text{Now, } (AC) \geq (AB) + (BC) - (B)$$

$$\Rightarrow (AC) \geq (A) + (B) - (B)$$

$$\Rightarrow (AC) \geq (A) \quad (I)$$

But (AC) can not be more than (A)

$$\text{i.e., } (AC) \leq A \quad (II)$$

From (I) and (II), it follows that

$$(A) = (AC) \quad \text{Proved.}$$

14.5 Joint Distribution of Attributes (Contingency Tables)

Let there be two attributes divided into two categories i.e. we are considering dichotomous classification attributes. In this case there are 9 class frequencies.

These cell frequencies together with their grand total give the joint distribution of the attributes because they show how the attributes vary jointly in the given group of individuals. From the joint distributions we also obtain two other types of distributions. Thus the row and column totals (marginal frequencies) together with the grand total gives the marginal distribution of the attributes.

These class frequencies in usual notations may be written in a 2×2 table as follows:

Attributes	A	α	Total
B	$(AB) = a$ (say)	$(\alpha B) = b$ (say)	(B)
β	$(A\beta) = c$ (say)	$(\alpha\beta) = d$ (say)	(β)
Total	(A)	(α)	N

This table is known as 2×2 contingency table.

Example.6: Let us consider an example of two dichotomous attributes Disease (Malaria) A and Precautionary measure (Vaccination Status) B in a certain population. Observed results (cell frequencies) are displayed in the following contingency table in different categories.

		Outcome (A)		Row Total
		Attacked with Malaria (A)	Not Attacked (α)	
Precautionary Measure (B)	Not Vaccinated (B)	(AB) 1	$(2)(\alpha B)$	$(B) = (1) + (2)$
	Vaccinated (β)	$(A\beta)$ (3)	$(4)(\alpha B)$	$(\beta) = (3) + (4)$
Column Total		$(A) = (1) + (3)$	$(\alpha) = (2) + (4)$	$(\beta) = (3) + (4)$

The cell frequencies shown in the table together with grand total give the joint distribution of the attributes under study. Thus the row total (Marginal frequencies) together with the grand total gives the distribution of the attribute “Precautionary measure” to be called the marginal

frequency distribution of precautionary measure. On the other hand column totals with N will give marginal distribution of “outcomes”. The other type of distribution is given by each column or row of frequency together with the corresponding column or row total of the table and it is said to give a conditional frequency distribution of outcome or precautionary measure of particular form. The contingency table need not be 2×2 only.

Check your progress

1. Define the attributes.
2. What do you mean by Consistency of Data?

14.6 Independence and Association of Attributes

Consider two dichotomous attributes A and B.

$$\text{So } (A) = (AB) + (A\beta)$$

$$(\alpha) = (\alpha B) + (\alpha\beta)$$

$$(B) = (AB) + (\alpha B) \quad (1)$$

$$(\beta) = (A\beta) + (\alpha\beta) \quad (2)$$

$$N = (AB) + (A\beta) + (\alpha B) + (\alpha\beta)$$

$$\text{Further, } N = (A) + (\alpha) \quad (3)$$

$$\text{And } N = (B) + (\beta) \quad (4)$$

Suppose that individual under consideration constitute the population itself and not just a sample from the population. Also ensure that none of the marginal frequencies is zero. Then the ratios $\frac{(AB)}{(A)}$ and $\frac{(\alpha\beta)}{(\alpha)}$ give, respectively, the proportions of members of population having B among those having (A) and among those having (α). If these proportions be equal we may say that the presence or absence of the character A in an individual does not in any presence or absence whether B will be present or not. A and B may then be called statistically independent or unrelated. Moreover, A and B are said to be associated if they are not independent.

For A and B to be independent we must have

$$\frac{(AB)}{(A)} = \frac{(\alpha\beta)}{(\alpha)} = \frac{(B)}{N}$$

Thus, $(AB) = \frac{(A).(B)}{(N)}$ then A and B are said to be independent.

If $(AB) > \frac{(A).(B)}{N}$ then A and B are said to be positively associated.

If $(AB) < \frac{(A).(B)}{N}$ then A and B are said to be negatively associated.

Remarks:

Two attributes are said to be completely associated if one of them cannot occur without the other though the other may occur without the one i.e. A's are B's are A's according to whether the A's or the B's are in minority. In other words,

If there is perfect positive association (complete) between A and B if all A's are B's and or all B's are A's i.e. if $(A\beta) = 0$ and/or $(\alpha B) = 0$

Similarly, complete dissociation may be defined as the case when no A's are B's no α 's are β 's. Or in other words.

There will be perfect negative association (complete) also known as complete dissociation if no A's are B's and/ or no α 's are β 's. i.e. $(AB) = 0$ and /or $(\alpha\beta) = 0$ Alternatively we may say that-

3. There is perfect absolute positive association between A and B if all A's are B's and all B's are A's i.e. if $(A\beta) = 0$ and $(\alpha\beta) = 0$
4. There is perfect absolute negative association between A and B if no A's are Bs and no B's are A's and no α 's are β 's. i.e. $(AB) = 0$ and $(\alpha\beta) = 0$.

Test of Independence of attributes (2×2 contingency table)

To test of independence of attributes (2×2 contingency table)

Attributes→ ↓	A		Row total
B	A	b	a+b
β	C	d	c+d
Column total	a+c	B+d	N= a+b+c+d

Where a,b,c and d are cell frequencies.

One uses the chi-square statistic defined as,

$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$ which follows chi-square distribution with one degrees of freedom under the assumption (null hypothesis) that the two attributes are independent.

If calculate value of chi-square χ^2 is greater than tabulation value of $\chi^2_1; \alpha$ then the null hypothesis that two attributes are independent is rejected otherwise the null hypothesis is accepted at $\alpha\%$ level of significance. Usually is taken as 5% or 1%. If the attributes come out to be associated we find the levels of association as follows.

14.7 Measures of Association for 2×2 cases

Let us define $\delta_{AB} = (AB) - \frac{(A).(B)}{(N)}$

1. Coefficient of association due to Yule

$$Q_{AB} = \frac{N\delta_{AB}}{(AB)(\alpha\beta) + (A\beta)(\alpha\beta)}$$

$$= \frac{(AB)(\alpha\beta) - (A\beta)(\alpha\beta)}{(AB)(\alpha\beta) + (A\beta)(\alpha\beta)}$$

Properties

- (i) $Q_{AB} = 0$ if $\delta_{AB} = 0$ i.e., A and B are independent
- (ii) $Q_{AB} = +1$ when $(AB)(\alpha\beta) = 0$ i.e. when $(AB) = 0$ and/or $(\alpha\beta) = 0$ i.e. when there is complete positive association between A and B.
- (iii) $Q_{AB} = -1$ when $(AB)(\alpha\beta) = 0$ i.e. when $(AB) = 0$ and/or $(\alpha\beta) = 0$ i.e. when there is complete negative association between A and B.

Coefficient of Colligation

Similar to Yule's coefficient of Association Q_{AB} having same general properties another measure Y_{AB} called Coefficient of Colligation. Also due to Yule is defined as follows:

$$Y_{AB} = \frac{\sqrt{(AB)(\alpha\beta)} - \sqrt{(A\beta)(\alpha\beta)}}{\sqrt{(AB)(\alpha\beta)} + \sqrt{(A\beta)(\alpha\beta)}}$$

Another measures of association is V_{AB} defined as, unlike

$V_{AB} = \frac{N\delta_{AB}}{(AB)(\alpha\beta) + (A\beta)(\alpha\beta)}$ has similar properties as Q_{AB} and Y_{AB} but Q_{AB} and Y_{AB} . $V_{AB} = \pm 1$ when and only when there is absolute association between two characteristics.

Remarks

1. If A and B are independent. A and β . α and B, α and β are also independent.
2. If A and B are negatively associated, then A and β are positively associated and α and β are negatively associated.

Manifold two-way (mxn) classification

Let us have two attributes where at least one occurs in more than two forms or subgroups. In general let two attributes A and B be divided in $A_1, A_2, \dots, A_m$ and $B_1, B_2, \dots, B_n$ forms or subgroups.

Further suppose cell frequency of $(I, j)^{th}$ cell be denoted by f_{ij} and express the number of individuals belonging to $(A_i, B_j)^{th}$ class or cell where $i = 1, 2, \dots, m$ and $j = 1, 2, \dots, n$. In this case mxn contingency table is represented as follows,

A \ B	B ₁	B ₂				B _n	
A ₁	f ₁₁	f ₁₂	---	f _{1j}	----	f _{1n}	(A ₁)
A ₂	F ₂₁	F ₂₂	-----	F _{2j}	----	F _{2n}	(A ₂)
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
A ₁	f _{i1}	f _{i2}	----	f _{ij}	---	f _{in}	(A _i)= f _{i0}
.							
A _m	f _{m1}	f _{m2}	-----	f _{m1}	-----	f _{mn}	(A _m)
Total	(B ₁)	(B ₂)		(B _j)= f _{0j}		(B _n)	$N = \sum_{i=1}^m \sum_{j=1}^n f_{ij}$

Marginal Frequency of A_i denoted as f_{i0} and marginal frequencies of B_j denoted as f_{0j} is given as follows

$$f_{i0} = \sum_{j=1}^n f_{ij}, \quad i = 1, 2, m$$

$$f_{0j} = \sum_{i=1}^m f_{ij}, \quad i = 1, 2, m$$

$$N = \sum_{i=1}^m f_{i0},$$

$$= \sum_{j=1}^m f_{0j}$$

$$\sum_{i=1}^n \sum_{j=1}^n f_{ij},$$

Assume that for $f_{i0} > 0$ and $f_{0j} > 0$ all i and j . In this case A and b are said to be unrelated or statistically independent if $f_{ij} = \frac{f_{i0}f_{0j}}{N}$ for all i and j . On the other hand if $f_{ij} \neq \frac{f_{i0}f_{0j}}{N}$ for any pair then A and B will be said to be associated.

$$= \sum_{j=1}^m f_{0j}$$

$$\sum_{i=1}^n \sum_{j=1}^n f_{ij},$$

Assume that for $f_{i0} > 0$ and $f_{0j} > 0$ all i and j . In this case A and b are said to be unrelated or statistically independent if $f_{ij} = \frac{f_{i0}f_{0j}}{N}$ for all i and j .

On the other hand if $f_{ij} \neq \frac{f_{i0}f_{0j}}{N}$ for any pair then A and B will be said to be associated.

Under the assumption that A and B are independent, contingency for (i, j)th cell is defined as

$$\delta_{ij} = f_{ij} - \frac{f_{i0}f_{0j}}{N}$$

$$\text{and } \chi_{AB}^2 = \sum_{i=1}^n \sum_{j=1}^n \frac{N\delta_{ij}^2}{f_{i0}f_{0j}}$$

$$= N = \sum_{i=1}^n \sum_{j=1}^n \frac{N\delta_{ij}^2}{f_{i0}f_{0j}} - N$$

Karl Pearson's Coefficient of Contingency $C_{AB} = \sqrt{\frac{\chi_{AB}^2}{\chi_{AB}^2 + N}}$

Tschuprow's Coefficient of contingency T_{AB} is defined as;

$$T_{AB} = \sqrt{\frac{\chi_{AB}^2}{N\sqrt{(m-1)(n-1)}}$$

$\chi^2_{AB} = 0$ if (if and only if) A and B are independent. This will also make values of C_{AB} and T_{AB} equal to zero. If there is a high association between A and B values of χ^2_{AB} , C_{AB} and T_{AB} will be high accordingly.

14.8 Yates Correlation

In a 2×2 contingency table for testing the independence of two attributes the degrees of freedom associated with chi-square statistic is $(2-1)(2-1) = 1$. Here for 2×2 contingency table we have $\chi^2 = \frac{N(ad-bc)^2}{(a+b)(c+d)(b+d)}$ follows chi-square distribution with one degree of freedom (d.f.)

If any of the cell frequencies less than five then use of pooling method for chi square results in chi-square with zero degrees of freedom (since one d.f. is lost in pooling) which is meaningless.

In this case we apply a correction due to F. Yates which is usually known as “Yates correction for continuity”. One should remember that chi-square is a continuous distribution and it fails to maintain its character of continuity if any one of the expected frequency is less than five. Because of this reason the name “correction for continuity” and it is applied only when one of the expected cell frequencies is less than five.

According to Yates correction we subtract (or add) $\frac{1}{2}$ from a and d and add (or subtract) $\frac{1}{2}$ to b and c so that marginal totals are not disturbed at all.

The corrected value of chi-square is given as

$$\chi^2 = \frac{N \left[\left(a \pm \frac{1}{2} \right) \left(d \pm \frac{1}{2} \right) - \left(b \pm \frac{1}{2} \right) \left(c \pm \frac{1}{2} \right) \right]^2}{(a+b)(c+d)(b+d)}$$

The degree of freedom is taken as unity.

Remarks

1. $\frac{1}{2}$ (or 0.5) is added to the cell frequency which is less than 5 and then adjusting for the remaining cell frequencies accordingly.
2. If N is large the use of Yates correction will make very little difference in the value of chi-square. However, if N is small the application of Yates correction may overstate the probability.
3. Yates correction has been widely recommended to every 2×2 table even if no theoretical cell frequency is less than five.

14.9 Exercises

E-1 Given that $(A) = (\alpha) = (B) = (\beta) = \frac{1}{2}$ show that $(AB) = (\alpha\beta)$.

E-2 Given that 50% of employees of a private intuition are men 60% are aged (over 50), 80% are highly qualified (post graduates), 35% aged men 45% highly qualified men and 42% highly qualified and aged find the greatest and least possible proportions of highly qualified aged men.

E-3 Is there any inconsistency in the following data:

$N=2200$, $(A)= 10000$, $(B) = 1400$ and $(AB)= 1200$?

Discuss with reasons.

E-4 In a series of houses actually invaded by TB 40% of the inhabitants are attacked and 65% have been vaccinated. What is the lowest percentage of the vaccinated that must have been attacked.

E-5 Investigate the association between darkness of eye colour in father and son from the following data:

Father with dark eyes and sons with dark eyes	=	60
Father with dark eyes and sons with non dark eyes	=	89
Father with not dark eyes and sons with dark eyes	=	99
Father with not dark eyes and sons with not dark eyes	=	792

Also tabulate for comparison the frequencies that would have been observed had there been no heredity.

14.10 Summary

Attribute is a qualitative characteristic which cannot be measured numerically. Simply we can count the presence or absence of a particular attribute in a group of individuals or unit. Sex, blindness, smoking, etc. are some examples of attributes.

If an attribute is divided in two categories, it is called dichotomy classification. However, if an attribute in more than two categories, it is called manifold classification.

In usual notations, two attributes A and B are said to be independent if $(AB) = \frac{(A)(B)}{N}$

If attribute are not independent they are said to be associated.

14.11 Further Readings

1. Goon, Gupta & Dasgupta: Fundamentals of Statistics Vol. I and Vol. II The World Press Pvt. Ltd., Kolkata.
2. Yule G.U. and Kerdall, M.G.: An Introduction to the Theory of Statistics, Charles Griffin and Company Ltd.
3. C.E. Weatherburn : Mathematical Statistics.