The Notes Project

Descriptive Statistics

December 9, 2023; rev. December 16, 2023

Keval Mehta, Divyansha Sachdeva

Part I

Descriptive Statistics - I

Syllabus

Unit 1: Types of Data and Data Condensation

1. Concept of population and sample. Different types of scales: nominal, ordinal, interval and ratio.
2. Collection of Primary data: concept of a questionnaire and a schedule, Secondary data
3. Types of data: Qualitative and quantitative data; Time series data and cross section data, discrete and continuous data.
4. Tabulation & Diagrammatic representation using bar diagrams, Line diagram and pie chart.
5. Univariate frequency distribution of discrete and continuous variables. Cumulative frequency distribution.
6. Graphical representation of frequency distribution by Histogram, frequency polygon, Stem and leaf diagram and Cumulative frequency curve.

Unit 2: Measures of Central Tendency

1. Concept of central tendency of data. Requirements of good measure
2. Mean, Median, Mode: Arithmetic mean (Simple, weighted mean, combined mean), Geometric mean, Harmonic mean, Median, Mode, Empirical relation between mean, median and mode
3. Partition Values: Quartiles, Deciles, Percentiles.
4. Merits and demerits of using different measures & its applicability

Unit 3: Measures of Dispersion, Skewness & Kurtosis

1. Concept of dispersion. Requirements of good measure.
2. Absolute and Relative measures of dispersion: Range, Quartile Deviation, Mean absolute deviation, Standard deviation.
3. Variance and Combined variance, raw moments and central moments and relations between them.
4. Concept of Skewness and Kurtosis: Measures of Skewness: Karl Pearson's, Bowley's and Coefficient of skewness based on moments.
5. Measure of Kurtosis
6. Box Plot

Unit 1: Types of Data and Data Condensation

What are statistics?

There are two subdivisions of statistical method:

- Descriptive Statistics - It deals with the presentation of numerical facts, or data, in either tables or graphs form, and with the methodology of analysing the data.
- Inferential Statistics - It involves techniques for making inferences about the whole population on the basis of observations obtained from samples.

Some Definitions

Population A population is the group from which data are to be collected. The entire group of individuals is called the population.

Sample Usually populations are so large that a researcher cannot examine the entire group. Therefore, a sample is selected to represent the population in a research study.

Data The measurements obtained in a research study are called the data. The goal of statistics is to help researchers organize and interpret the data.

Types of Statistical data

- Primary data is data collected primarily for the purpose of the given enquiry. They are original in character and are raw materials of the enquiry.
- Secondary data are data already collected by someone's for some purpose and are available for the present study.

Primary Data Collection Methods

Direct personal observation

- In this method, the investigator collects the data personally.
- They ask or cross-examine the informant in a tactful and courteous manner, and collect the necessary information.
- This method is adopted when
  - intensive or in-depth study is essential.
  - greater accuracy is needed.
  - the field of enquiry is not large but complex.

  - data of a confidential nature are to be collected.
  - sufficient time is available.
- Such a personal enquiry gives reliable and accurate information as response will be encouraging because of personal touch.
- Also uniformity and homogeneity of data can be maintained.
- However, such an enquiry will be expensive and time consuming & will not give good results at the hands of an untrained investigator.
- Also, the chances of personal prejudices/ bias creeping into the investigation are more.

Indirect oral investigation

- This method is used when the informant is reluctant to supply information.
- Here the investigator approaches witnesses or third parties who are in touch with the informant. For instance if one wants to collect information about drinking or gambling habits of people one gets the information from family members, friends and/or liquor shops, etc.
- This method is generally adopted by Government agencies for their enquiry committees of commissions.
- Also, in cases of thefts, murders, riots etc. The police interrogate third parties who possess knowledge about the happenings under study.
- This method is simple and convenient, and adequate information can be obtained if information is collected from different parties.
- However, absence of direct contact can mar the reliability of the information.
- Also, witnesses may colour the information to suit their interests.

Data through Mailed Questionnaires/Online Data Collection

A. Mailed Questionnaire

- As the title suggests data is collected through a questionnaire consisting of a list of questions pertaining to the enquiry.
- This questionnaire is posted to the respondents, who are expected to answer the questions or write the answers in the blank spaces. A covering letter is also sent along with the questionnaire requesting full co-operation and prompt reply from the respondents
- This method is followed by research workers, private individuals, non-official agencies and Government agencies. The mailed questionnaire method is the most economical and there is a saving of time and labour. It is suitable when the area of the survey is large.
- Since information is obtained directly from respondents error in the investigation is small.
- However, since there is no direct contact between investigator and respondent one can not be sure about the accuracy and reliability of the data.
- Some people may not reply as they may be illiterate or simply lazy.
- This could lead to non-response or delayed response.

B. Online Data Collection

- As the title suggests data is collected through a questionnaire consisting of a list of questions pertaining to the enquiry through an online method using internet.
- This questionnaire/form is posted to the respondents, who are expected to answer the questions or write the answers in the blank spaces. A covering message is also sent along with the questionnaire requesting full co-operation and prompt reply from the respondents.
- The methods used are Google Forms, Computer-Assisted Telephone Interviewing (CATI) etc.

Information through agencies

- Here, the investigator appoints agents or correspondents to collect data. These agents collect the information and transfer it to the investigator.
- This method of primary data collection is usually followed by news agencies where information is needed in different fields like politics, sports, natural and man-made calamities etc.
- This method is adopted where information is required from a wide area on a regular basis.
- Although this method gives extensive information in a speedy and economical way, it may be biased and the requisite degree of accuracy and uniformity can not be maintained.

Investigation through enumerators

- This method is generally employed by the Government for popul ation census etc.
- A number of enumerators are selected and trained. They are then armed with standardised questionnaires and sent to informants to get first hand information.
- This method is adopted when one requires reliable and accurate information from all types of people (literate and illiterate).
- However, this data collection exercise may be costly and time consuming, and depends on the efficiency of the enumerators.

Questionnaires

From the above primary data collection methods, the most important requirement for a successful data collection exercise is a questionnaire. There are two sections in any questionnaire. They are:

Classification This section includes the details of the respondents such as name, age, sex, education, marital status, occupation. Additionally, details like the date of interview and name of the interviewer are included in this section.

Subject Matter This section includes questions related to the subject matter of the inquiry. The answers given here can be analysed according to the information in the classification section

Requisites of a Good Questionnaire

- The questionnaire must be brief, i.e. the number of questions should be as few as possible as respondents will not like or may not have time for answering a long questionnaire. All the questions must be relevant to the problem under investigation
- Questions should not be ambiguous. They must be capable of one and only one interpretation.
- Questions must be easily understood. Technical terms must be avoided except when addressed to specialists.
- Questions must be arranged in a logical sequence
- Questions should have a precise answers. The answers should take the form of ‘yes’ or ‘no’, a quantity, a date, a place etc. Wherever possible the questionnaire should suggest answers so that the informant has merely to tick or cross (✓ or ✗) the answer.
- Questions must not contain words of vague meaning. To ask if something is large or if a man is unskilled are examples of such questions.
- Questions of a sensitive or personal nature should be avoided. Such questions may not be answered and some informants may be offended.

- Questions should not require the respondent to make any calculations. The figures provided by the respondent must be accepted and calculations according to the need of the questionnaire done later.
- To check reliability of answers. Some questions should be asked so as to provide cross-check for the answers to the similar questions.

Schedules

- This method of data collection is very much like the collection of data through questionnaire, with little difference which lies in the fact that schedules (proforma containing the set of questions) are filled in by the enumerators who are specially appointed for the purpose.
- These enumerators go to respondents, put to them the questions from the schedule and record the replies in the space meant for the same in the proforma.
- The method require careful selection of enumerators for filling up schedules and they should be trained to perform their job well. The enumerators should be honest, sincere, hardworking and should have patience and perseverance.
- This method of data collection is very useful in extensive enquiries and can lead to fairly reliable results.
- It is, however very expensive.
- Population census all over the world is conducted through this method.

Differences Between Questionnaires and Schedules

Questionnaire	Schedule
Sent through mail to informants with a covering letter.	Filled in by the enumerator
Relatively **cheap** and economical, since we have to spend money only in preparing the questionnaire and mailing it to the respondents. No field staff is required.	Relatively **more expensive** since a considerable amount of money is spent in appointing enumerators and training them.
Non-response is high in case of questionnaire as many people do not respond, and many return questionnaire without answering all questions.	Non-response is generally very low in case of schedules.
The questionnaire method is likely to be **slow** since many respondents may not return the questionnaire on time.	In case of schedule the information is **collected well in time** as they are filled in by the enumerators.
Questionnaire method can be used only when the respondents are literates and co-operative.	in case of schedules, the information can be gathered even when the respondents happen to be illiterate.
It is not clear as to who replies.	The identity of respondent is known.

Editing the Data

- Once the data has been obtained either from primary or a secondary source, the next step in any statistical investigation is to edit data.
- The main purpose of editing the data is to scrutinize it for possible errors and irregularities.
- The task of editing is a highly specialised one and necessary too before proceeding to tabulation.
- It requires great care, and negligence in this regard render a valuable study useless.

The Process

- Check if the schedules/questionnaire for completeness: The editor must see whether answer to each question has been furnished. If not, the informant must be contacted again. If one doesn't still get any reply then editor should mark 'no reply' in space provided for the answer.
- Ensuring that there are no inconsistent entries: The editor must check if answers to questions are not contradictory.
- For example, if answers to two questions 'are you married?' and 'how many children do you have?' are 'no' and 'two' respectively, then clarification should be used.
- Checking the questionnaires for accuracy: Validity or reliability of conclusions drawn from survey depends on accuracy of information. e.g. shifting of a decimal point. Also it is obvious that son's age can not be 20 if the mother's age is reported as 25. Entries reported for Sep 31 or Feb 30 cannot be considered.
- Checking the questionnaire for homogeneity of answers: The editor must check if the questions have been interpreted uniformly.
- e.g. Gross pay vs. net pay or monthly pay vs. annual pay.

Types of Data

The data can be of the following types:

1 – Geographical or Spatial
– Chronological or Time-Series

2 – Qualitative
– Quantitative

3 – Discrete
– Continuous

Measuring Variables

- To establish relationships between variables, researchers must observe the variables and record their observations. This requires that the variables be **measured**.
- The process of measuring a variable requires a set of categories called a **scale of measurement** and a process that classifies each individual into one category.

Nominal

- A nominal scale is an unordered set of categories identified only by name.
- Measurement is done only in terms of whether the individual terms belong to some distinctively different categories
- We cannot quantify or even rank order these categories.
- Each category can be assigned a number (group 1, group 2 etc.). The assignment of numbers to qualitative classes is purely arbitrary since any number would do as well as another
- The characteristic measured on nominal scale generates categorical data.

e.g. Data on gender, race, colour, city, etc.

Ordinal

- An ordinal scale is an ordered set of categories.
- This scale allows us to rank the items according to the classes but the differences between the adjacent ranks may not be equal.

- It can be said that one class is higher than the other.
- Ordinal measurements tell you the direction of difference between two individuals.
- Nominal measurement provides less information than ordinal measurements.

e.g. Socioeconomic status of families: Upper Class, Middle Class and Lower Class

Interval

- An interval scale is an ordered series of equal-sized categories.
- Interval variables allow us not only to rank order the items that are measured but also to quantify and capture the sizes of differences between them.
- There is no true(absolute) zero.
- The Fahrenheit scale is an example of interval scale.
- An increase in temperature from 30° to 40° involves the same increase in temperature from 60° to 70°. Equal temperature differences are equal in the sense that the same amount of heat is required to raise the temperature of an object from 30° to 40° , or 60° to 70° . But $^{60°}/_{30°} = 2$ need not imply that 60° is twice as hot as 30° i.e. the ratio has no significance.
- Unlike nominal and ordinal, it is a truly quantitative scale.
- There is a constant (equal) size interval between any adjacent units on the measurement scale (unit distance) e.g. the difference between 40° and 30° is same as that between 70° & 60° .
- the existence of zero but does not necessarily indicate absence of quantity being measured (0° Fahrenheit does not necessarily mean absence of heat).
- Some interval scales in biological data collection are circular scales (time of the day or time of a year where zero is arbitrary).
- There is no inherent starting point and ratios are meaningless.

e.g. temperatures, time or pressure

Ratio

- A ratio scale features an identifiable absolute zero point
- It is an interval scale where a value of zero indicates none of the variable.
- Ratio measurements identify the direction and magnitude of differences and allow ratio comparisons of measurements.
- Most statistical data analysis procedures do not distinguish between the interval and ratio properties of the measurement scales.
- In a ratio scale, both differences and ratios are meaningful.
- There is a constant (equal) size interval between any adjacent units on the measurement scale (unit distance)
- There is a True Zero Point. This enables us to say something about the ratio of measurements. E.g. A person weighing 100 kgs weighs double another weighing 50 kgs.

e.g. weight, height or distance

Analyses for Different Scales

- Statistical analysis of the data depends on the scale of measurement.
- For nominal scale, mode is the appropriate measure of central tendency and the χ^2 test is used for statistical significance.
- In case of ordinal scale, the appropriate measures of central tendency and dispersion are median and quartile deviation. The tests of statistical significance are restricted to non-parametric tests.
- For interval and ratio scales, the appropriate measures of central tendency and dispersion are mean and standard deviation respectively. The tests for statistical significance are t-test, F-test etc.

Classification of Data

Why?

- To condense raw data into a compact form which is easily understood and can be quickly interpreted.
- To bring uniformity in large mass of data.
- To facilitate comparison of two or more characteristics presented.

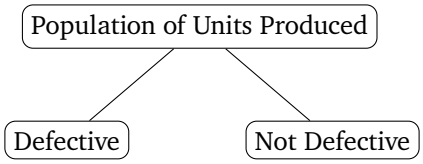

Figure 1: One-Way Classification

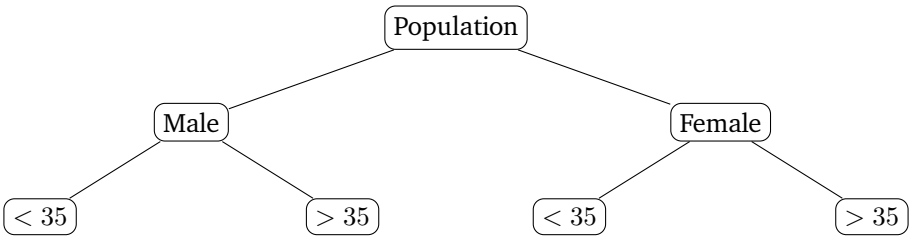

Figure 2: Manifold Classification

Tabulation

Tabulation may be defined as a logical and systematic arrangement of statistical data in rows and columns.

Why?

- To simplify complex data
- To clarify the aim of the survey
- To do away with unnecessary details and save space
- To make comparisons easy
- To facilitate statistical analysis
- To facilitate future studies in same or relevant field
- To identify trends or patterns of data

A statistical table is a systematic presentation of numerical data in rows and columns with respect to some characteristics of the population.

Requirements of a good statistical table

Title Each table must have a title displayed in bold type at the top of the table. It must be brief and self-explanatory.

Department	No. of Employees
R&D	9
HR	13
Marketing	19
Finance	12
Sales	7
Total	60

Table 1: A Simple Table

Department	No. of Employees		
	Male	Female	Total
R&D	6	3	9
HR	8	5	13
Marketing	10	9	19
Finance	8	4	12
Sales	4	3	7
Total	36	24	60

Table 2: A Complex Table (Two-Way)

Number For easy reference of information in the table, it must be numbered. The number can be put at the top above the title or at the bottom.

Stubs and Captions Stubs are row headings and captions are the column headings. These headings should be well-defined and brief.

Body It is the most important part of any table. It contains numerical information including meaningful row and column totals.

Source-Note It mentions the source/origin of information used in table.

Footnote Normally placed at the bottom of table, it gives explanation or elaboration about any information used in table.

Simple and Complex Tables

- Simple and Complex tables are tables used to represent one or more characteristics under study.
- In simple table the data are classified according to only one characteristic.
- In a complex table two or more characteristics are shown.

Classifying statistical tables

General Purpose

- General purpose tables are tables published by the Government and International bodies.
- They are used as reference tables and they provide information about facts.
- They are constructed to keep a record of the data collected as in census.

Special Purpose

- Special purpose tables, also called summary tables are constructed to suit a particular purpose.
- They are analytical in nature or derived from general purpose tables.

Department	No. of Employees								Grand Total
	Male			Female			Total		
	< 35	> 35	Total	< 35	> 35	Total	< 35	> 35	
R&D	3	3	6	2	1	3	5	4	9
HR	6	2	8	2	3	5	8	5	13
Marketing	7	3	10	6	3	9	13	6	19
Finance	5	3	8	3	1	4	8	4	12
Sales	4	0	4	2	1	3	6	1	7
Total	25	11	36	15	9	24	40	20	60

Table 3: A Complex Table (Three-Way)

Diagrammatical Presentation of Data

Why?

- Besides tabulation, data can also be understood better if it was presented with the help of a diagram or a graph.
- Cold figures may not be very inspiring to people but diagrams help us see the pattern and shape of the data.
- Diagrams are a visual device to present data and they appeal to a layman too.

We use the following:

One Dimensional

- Bar Diagram
  - Simple
    - A simple bar diagram is a bar diagram used to represent one variable
    - For example – figures of sales, population (for various years, countries)
    - Simple bar diagram is the most common type of diagram used in practice
    - This diagram consists of vertical/horizontal bars of uniform width and height proportional to the value of the variable which the diagram seeks to present
    - It is called one-dimensional as it is only the height of the bar that matters and not the width
    - The space between the neighbouring bars must be uniform
    - Bars can be vertical or horizontal, but vertical bars are preferred as they are easier to comprehend and convenient for comparison
  - Subdivided
    - This diagram is used where a variable is divided into different components and changes in the values of these components as well change in the value of a variable are important
    - Each bar gives a comparative study
    - The bar is subdivided into different components
    - Each component occupies a part of the bar proportional to its share in the total
    - To distinguish different components from one another different colours or patterns are used
    - Here a key to the diagram is absolutely essential
    - The subdivisions of the bar are either in absolute figures or as percentages of the whole
    - If percentages are used, then the diagram is called a percentage sub-divided bar diagram
    - Here, all the bars are of the same height
  - Multiple

    - Multiple bar diagram is used to represent two or more sets of values which are related
    - The bars of the different sets are drawn adjacent to each other and the heights of the bars are proportional to the values
    - To distinguish between adjacent bars, different colours or patterns are used
    - Here, a key to the diagram is essential
    - This diagram can be used to represent different variable or one variable when the variable is divided into different components and the changes in the value of these components separately are important rather than the change in value of variable as a whole
- Line Diagram
  - A style of chart that is created by connecting a series of data points together with a line
  - This is the most basic type of chart used in finance and it is generally created by connecting a series of past prices together with a line

Two Dimensional

- Pie Diagram
  - A pie diagram is a circle which represents the total value of a variable
  - The circle is divided into a number of sectors representing different components of the variable
  - These sectors are such that their areas are proportional to the values of the components i.e. the angles of the sectors are taken proportional to the values of the components
  - Angle of each sector $= \frac{\text{Value of the component}}{\text{Total}} \times 360$
  - The radius of the circle is taken proportional to the square-root of the total
  - The diagram is so called because it looks like a pie and the components resemble slices cut from it
- Square Diagram
- Rectangular Diagram

Other

- Stem and Leaf Diagram

A Few Bar, Line, and Pie Graphs

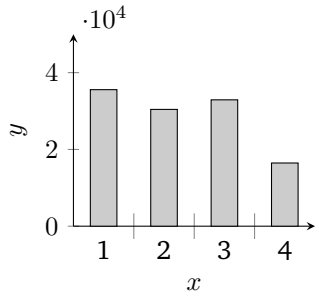

Figure 3: Simple Bar

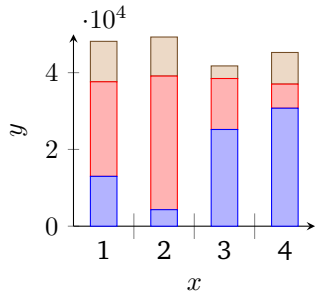

Figure 4: Subdivided Bar

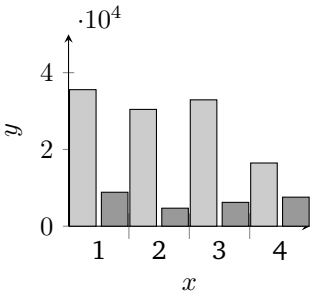

Figure 5: Multiple Bar

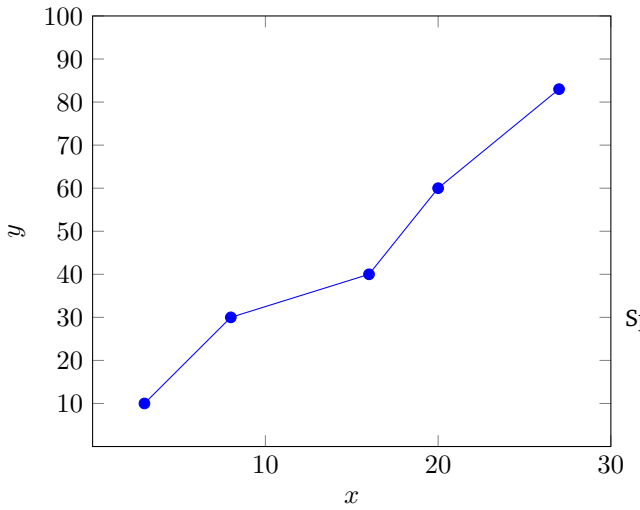

Figure 6: Line

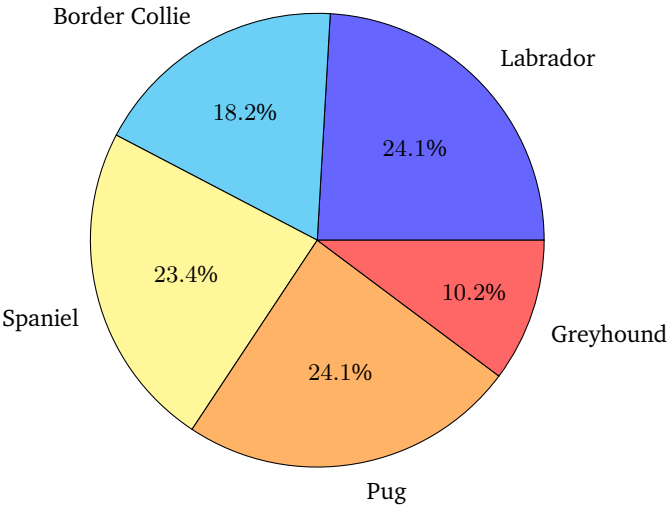

Figure 7: Pie

Advantages and Limitations

Advantages

- It is easily understood by all including people who have no background of statistics
- Data can be presented in a more attractive form and is appealing to the eye
- Comparative analysis and interpretation is possible

Limitations

- Diagrams do not show all the facts in detail
- It is a supplement to tabular representation but not an alternative to it
- The uses of certain diagrams are limited to experts
- If there is a wide gap between two different measurement, diagrams will not give a meaningful look

Frequency Distributions

- The collected data (after editing) is called raw data
- The raw data is in an ungrouped form i.e. the observations on the variate are just a set of numbers
- We need to put them in the form of a table for better understanding
- Such a table consists of two columns:
  - classes
  - class frequency,

  which denotes number of observations belonging to each class.
- Classes can be of three kinds:
  - a discrete set of values

  e.g. $0, 1, 2, \ldots$
  - a discontinuous (inclusive) set of class intervals

  e.g. $0 - 9, 10 - 19, \ldots$
  - a continuous (exclusive) set of class intervals

e.g. $0-10, 10-20, \ldots$

- Sometimes, for statistical analysis, exclusive type of class intervals are required.
- Hence, if the given frequency table is of inclusive type, it can be converted into exclusive type as follows:
  1. Find the difference between the lower limit of a class and the upper limit of an earlier class
  2. Subtract half of the difference from every lower class limit and add it to every upper class limit

Example Calculation

Question: The following data is about rainfall (in mms) in the month of July in a certain place. Prepare a frequency table taking class intervals $(40-50), (50-60), \ldots$

57.6, 72.8, 48.1, 71.4, 83.1, 91.6, 71.3, 63.4, 43.9, 69.2, 87.5, 90.1, 98.8, 49.2, 54.6, 71.5, 62.7, 59.7, 48.3, 54.1, 73.6, 48.2, 54.6, 77.1, 49.6, 58.3, 60.5, 63.2, 54.7, 65.0, 70.1

Solution: The table is formed as below:

Rainfall (in mms)	Tally Marks	No. of Days
40-50	𝍸 |	6
50-60	𝍸 ||	7
60-70	𝍸 |	6
70-80	𝍸 ||	7
80-90	||	2
90-100	|||	3
	Total	31

Table 4: Frequency Distribution Table

- The tally marks column is essential for obtaining the class frequencies by reading the data
- While calculating frequency, each observation is read and a tally is put against the corresponding class
- Every fifth tally mark is represented by putting cross tally
- This technique facilitates the counting of tally marks
- While constructing a frequency table using discontinuous or continuous class intervals, we need to decide:
  1. How many classes would be ideal?
  2. What should be the width of each of these classes?
- For the answer to the first question, we can use the formula suggested by **Struges**:

$$k = 1 + 3.322 \log_{10} N$$

- where, $k =$ number of classes, and
- $N =$ number of observations.
- For the answer to the second question, we can use the formula also suggested by **Struges**:

$$\text{Width} = \frac{\text{Maximum Observation} - \text{Minimum Observation}}{\text{Number of Classes}}$$

Graphical Presentation

A graphical representation of the frequency table would be ideal for even a layman to understand, as compared to a frequency distribution. Various ways of depicting a frequency distribution on a graph are:

Frequency Curve/Polygon

- This graphical presentation of a frequency distribution uses class marks and frequencies of the corresponding classes
- To construct a frequency curve we take the value of observed variable x on the $X-$ axis and the frequency (f) along the $Y-$axis
- Thus, we plot points (x, f) on the graph
- Then, we join these points by a smooth curve to obtain a frequency curve for a discrete frequency distribution
- If the frequency distribution is in the form of class intervals then x will be the class mark of the class interval
- For a frequency polygon, the points are joined by straight lines instead of a smooth curve and the end points are joined to the $X-$axis

Histogram

- One of the most important and useful methods of presenting a frequency distribution of a continuous variable (i.e. continuous class- intervals) is a **Histogram**
- A histogram is a graph containing a set of adjacent rectangles with width (on the $x-$axis) equal to each class interval and corresponding frequency (on $y-$axis) being the height of the rectangle, provided the class intervals are of the same width
- If the class intervals are of unequal width, height of the rectangle is taken as frequency density of each class interval, where

$$\text{Frequency Density} = \frac{\text{Frequency}}{\text{Width}}$$

Ogives

- Ogives are one more way of presenting a continuous frequency distribution
- Ogives or cumulative frequency curves are of two types
  - Less than $(<)$ type ogive
    - While drawing less than type of ogive, upper class boundaries are taken along the $X-$axis and less than cumulative frequency on Y-axis
    - Less than cumulative frequency of the first (lowest) class interval is frequency of that class interval
    - Less than cumulative frequency of the remaining class intervals are calculated by adding cumulative frequency of the previous class interval to the frequency of the class interval
    - Less than cumulative frequency represents the number of observations less than the upper class boundary of the class interval
  - Greater than or equal to $(\geq)$ type ogive
    - While drawing greater than or equal to type ogive, lower class boundaries are taken along the $X-$axis and greater than or equal to cumulative frequency on $Y-$axis.
    - More than or equal to cumulative frequency of the first (lowest) class is total frequency.
    - For the remaining class intervals they are calculated by subtracting frequency of the earlier class interval from the cumulative frequency of the earlier class interval.
    - More than or equal to cumulative frequencies represent the number of observations greater than or equal to the lower class boundary of the class interval
  - In both the above cases, the points plotted on the graph are joined by a smooth curve called an **ogive**

Stem and Leaf Diagram

- Stem and leaf diagrams represent ungrouped data
- Ungrouped data are also represented by a frequency table, but once we have the frequency table we lose some information on how the observations are distributed in a class interval
- This loss of information is avoided in stem and leaf displays
- Also a stem and leaf display gives us the rank order of the item in the data set, and the shape of the distribution too
- It is a very useful technique in exploratory data analysis

Steps

1. Divide each value of the observation into two parts. One part consists of one or more leading digits as stem and the rest of the digits as the leaf
2. For a 3 digit no. the first 2 digits form the stem and last digit forms the leaf
3. The stem values are listed out to the left of a vertical line and each leaf value corresponding to a stem is written in a horizontal line to the right of the stem in the order in which they are encountered in passing from one observation to the other
4. Finally, we arrange all the leaves in each row in ascending order

- The stem and leaf diagram serves the same purpose as a histogram i.e. to provide a visual impression of the distribution
- A histogram shows only the number of observations in each class-interval while a stem and leaf diagram shows the actual observations

Less Than Ogive, Histogram, and Stem and Leaf Diagrams

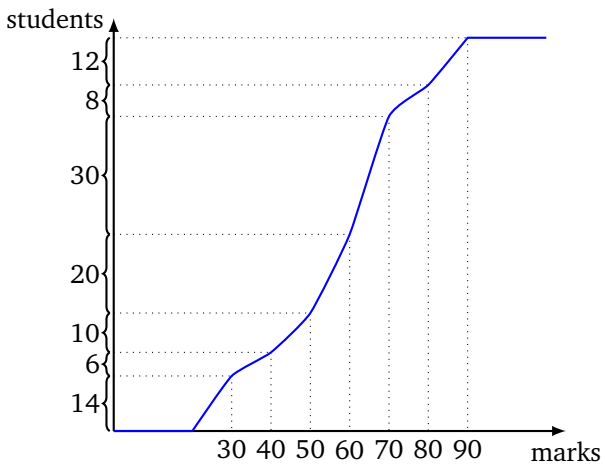

Figure 8: Ogive

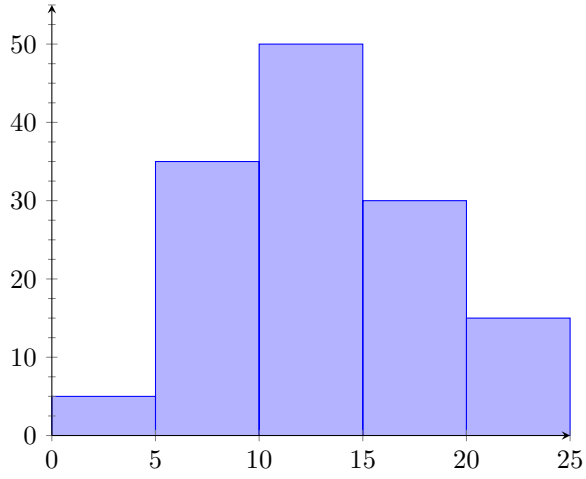

Figure 9: Histogram

For data given to be: 57,94, 84, 63, 72, 51, 50, 48, 61, 71, 93, 82, 100, 89, 67, 78, 79, 78, 76, 85, 87, 73, 66, 99, 84, 72, 66

The stem-and-leaf diagram is:

Stem	Leaf							
4	8							
5	7	1	0					
6	3	1	7	6	6			
7	2	1	8	9	8	6	3	2
8	4	2	9	5	7	4		
9	4	3	9					
10	0							

Table 5: Stem-and-Leaf Diagram

Unit 2: Measures of Central Tendency

What are they?

- It is sometimes essential to condense the data into a single value which can describe the characteristics of the entire data
- This single value is referred to as an average or central value or measure of central tendency
- In practice, the word 'average' is used with different meanings
- In some cases, we use term 'average' to denote a mediocre type e.g. average student, average film, average actor etc.
- In some other cases, by 'average' we mean 'typical' or 'usual' e.g. average height of an Indian, average income etc.
- In statistics, 'average' refers to a value obtained by a specific process e.g. average height, average income etc.
- In statistical terms, average is a value around which most of the observations in the data are clustered
- It usually lies somewhere near the center of the group of observations and hence averages are termed as measures of central tendency
- This single value depicts the main characteristics of the data
- Large volumes of data cannot be easily understood or remembered
- So, a single value summarizing prominent features of the data is needed
- If two or more sets of data are to be compared, it is not possible to compare each and every item
- So, we require a single value representing the entire data in condensed form
- Thus, the objectives of a measure of central tendency or an average are:
  - To obtain a single representative quantity for the entire data
  - To facilitate comparison
- The most commonly used measures of central tendency are:
  1. Mean(s):
     a. Arithmetic
     b. Geometric
     c. Harmonic
  2. Median
  3. Mode
- The means are mathematical averages, while the median is a positional average

What makes a good average?

- It should be simple to understand and easy to calculate
- It should be rigidly defined
- It should be based on all observations in the data and sensitive to changes
- It should be capable of further mathematical treatment
- It should be least affected by extreme observations or sampling fluctuations
- It should be a true representative of the data

Properties, Merits and Demerits

Arithmetic Mean

Properties

- If a constant is added or subtracted from each observations in the data, the A.M. of the data will be affected in the same way e.g. if each observation in the data is multiplied by 2, the A.M. of the new data will be two times the original A.M.

Merits

- It is rigidly defined
- It is easy to understand and easy to calculate
- It is based on all the observations
- It is capable of further algebraic treatment
- Of all the averages A.M. is least affected by sampling fluctuations i.e., it is a stable average

Demerits

- It cannot be obtained by mere inspection nor can it be located graphically
- It cannot be obtained even if a single observation is missing
- It is affected by extreme values
- It cannot be calculated for frequency distribution having open end class intervals e.g. class-intervals like below 10, above 50, etc.
- It may be a value which may not be present in the data
- Sometimes, it gives absurd results e.g. average number of children per family is 1.28
- It cannot be used for the study of qualitative data such as intelligence, honesty, beauty, etc.

Even though A.M. has various demerits, it is considered to be the best of all averages as it satisfies most of the requisites of a good average. It is called the **ideal average**.

Median

Properties

- Median is the value of the central observation in the data when the observations are arranged in increasing (or decreasing) order of their magnitude.
- It divides the data into two parts containing equal number of observations above and below it.

Merits

- It is easy to understand and easy to calculate
- It is quite rigidly defined
- It can be computed for a distribution with open-end classes.
- In majority of the cases, it is one of the values in the data
- It can be determined graphically
- Since median is a positional average, it can be computed even if the observations at the extremes are unknown
- It is not highly affected by fluctuations in sampling

- It can be calculated even for qualitative data.

Demerits

- When the number of observations is large, the pre-requisite of arranging observations in ascending/descending order of magnitude is a difficult process.
- It is not based on all observations and hence, may not be a proper representative.
- It is not capable of further mathematical treatment.
- Since it does not require information about all observations, it is insensitive to some changes.

Mode

Properties

- Mode is defined as the value of the variable which occurs most frequently in the data
- It is a value which is repeated maximum number of times or with high frequency
- Thus, mode is considered as the most typical value of the data

Merits

- It is easy to understand and simple to calculate.
- It is not affected by extreme values or sampling fluctuations
- It can be calculated for distribution with open-ended classes.
- It can be determined graphically.
- It is always present within the data and is most typical value of the given set of data
- It is applicable to both qualitative and quantitative data

Demerits

- It is not rigidly defined.
- It is not based on all observations
- It is not capable of further mathematical treatment
- It is indeterminate if the modal class is at the extreme of the distribution
- If the sample of data for which mode is obtained is small, then such mode has no significance.

Calculating these measures

Arithmetic Mean

Raw Data

$$\overline{x} = \frac{\text{Sum of Observations}}{\text{Number of Observations}} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Ungrouped Frequency Distribution

$$\overline{x} = \frac{f_1 \cdot x_1 + f_2 \cdot x_2 + \cdots + f_n \cdot x_n}{f_1 + f_2 + \cdot + f_n} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

where f_i's are the frequencies of each individual x_i's.

Grouped Frequency Distribution

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

where x_i's are the class marks of each individual class.

Combined

$$\overline{x} = \frac{\sum_{i=1}^{k} n_i \overline{x_i}}{\sum_{i=1}^{k} n_i}$$

where n_i's are the population sizes of k populations, and $\overline{x_i}$'s are their respective population arithmetic means.

Weighted

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

where w_i's are the weights of each individual x_i's.

Geometric Mean

Raw Data

$$G = \text{antilog}\left[\frac{\sum_{i=1}^{n} \log x_i}{n}\right]$$

Frequency Distribution

$$G = \text{antilog}\left[\frac{\sum_{i=1}^{n} f_i \log x_i}{n}\right]$$

where f_i's are the frequencies of each individual x_i's, and x_i's are the class marks of each individual class in case of grouped frequencies.

Harmonic Mean

Raw Data

$$H = \frac{n}{\sum_{i=1}^{n} (1/x_i)}$$

Frequency Distribution

$$H = \frac{\sum_{i=1}^{n} f_i}{\sum_{i=1}^{n} (f_i/x_i)}$$

where f_i's are the frequencies of each individual x_i's, and x_i's are the class marks of each individual class in case of grouped frequencies.

Let a and b be two positive numbers, and A, G, and H be their arithmetic, geometric, and harmonic means respectively.

$$\begin{aligned}
A &= \frac{a+b}{2} \\
G &= \sqrt{ab} \\
H &= \frac{2}{1/a + 1/b} = \frac{2ab}{a+b} \\
A \cdot H &= \frac{a+b}{2} \cdot \frac{2ab}{a+b} \\
&= ab \\
&= G^2 \\
\implies G &= \sqrt{AH}
\end{aligned}$$

Next,

$$\begin{aligned}
A - G &= \frac{a+b}{2} - \sqrt{ab} \\
&= \frac{1}{2}(a - 2\sqrt{ab} + b) \\
&= \frac{1}{2}(\sqrt{a} - \sqrt{b})^2 \\
&\geq 0 \\
\implies A &\geq G
\end{aligned} \tag{i}$$

Also,

$$\begin{aligned} G - H &= \sqrt{ab} - \frac{2ab}{a+b} \\ &= \frac{\sqrt{ab}}{a+b}(a - 2\sqrt{ab} + b) \\ &= \frac{\sqrt{ab}}{a+b}(\sqrt{a} - \sqrt{b})^2 \\ &\geq 0 \\ \implies G &\geq H \end{aligned} \tag{ii}$$

From (i), (ii),

$$A \geq G \geq H$$

Median

Raw Data

n **is odd** The median is the $\left(\frac{n+1}{2}\right)^{\text{th}}$ observation.

n **is even** The median is the average of the $\left(\frac{n}{2}\right)^{\text{th}}$ and $\left(\frac{n}{2}+1\right)^{\text{th}}$ observation.

Ungrouped Frequency Data

Step 1. Find less than type cumulative frequencies for the given ungrouped frequency distribution.

Step 2. Find the total of the frequencies, say $N = \sum f$.

Step 3. If N is odd, find the value of the $\left(\frac{n+1}{2}\right)^{\text{th}}$ observation. The median is that value of the variable at which the less than type cumulative frequency is equal to or exceeds $\left(\frac{n+1}{2}\right)$ for the first time.

Step 4. If N is even, find the value of the variable at which the less than type cumulative frequency is equal to or exceeds $\left(\frac{n}{2}\right)$ and $\left(\frac{n}{2}+1\right)$ for the first time. If these two values are same, this value is the median of the distribution. If the two values are different, the median is their arithmetic mean.

Grouped Frequency Data

Step 1. Find less than type cumulative frequencies.

Step 2. Locate the median class (Median class is the class-interval in which the value of the median i.e. where the value of $\frac{n}{2}^{\text{th}}$ observation falls). It is that class-interval where less than type cumulative frequency is equal to or exceeds $\frac{n}{2}$ for the first time.

Step 3. The value of median (M) can then be calculated using the formula

$$M = l_1 + \frac{(l_2 - l_1)(\mathrm{N}/2 - \text{c.f.})}{\mathrm{f}}$$

where,

l_1 lower class-boundary of the median class

l_2 upper class-boundary of the median class

$\mathrm{c.f.}$ less than type cumulative frequency of the pre-median class (i.e. the class interval just preceding to the median class)

f frequency of the median class

N total frequency

Mode

Raw Data and Ungrouped Frequency Distribution

- If the data is available in the form of individual observations, mode can be obtained by inspection or in case of large number of values after converting the data as an ungrouped frequency distribution.
- For ungrouped frequency distribution, the mode can be determined merely by inspection as the value of the variable having the maximum frequency.

Grouped Frequency Distribution

Step 1. Suppose the data is represented in the form of a continuous frequency distribution as class-intervals and corresponding frequencies. (If the given class-intervals are inclusive type, they must be converted to continuous or exclusive type before calculating the modal value.)

Step 2. Then, locate the class-interval corresponding to the highest frequency. This class-interval is called as the modal class.

Step 3. The mode can then be calculated by the following formula:

$$M = l_1 + \frac{(l_2 - l_1)(f_1 - f_0)}{(f_1 - f_0) + (f_2 - f_1)}$$

where,

l_1 lower class-boundary of the modal class
l_2 upper class-boundary of the modal class
f_1 frequency of the modal class
f_0 frequency of the pre-modal class
f_2 frequency of the post-modal class

Step 4. The assumptions for this method are:

i. The frequency distribution must be continuous with exclusive type classes and having unique maximum frequency.
ii. The width of all the class-intervals must be the same. In case of distribution with unequal class-intervals, they should be made equal under the assumption that frequencies are uniformly distributed over the classes.
iii. Mode cannot be determined if the modal class is at the extreme i.e. the maximum frequency occurs at the beginning or at the end of the frequency distribution. To overcome this difficulty in computing mode, the empirical relationship Mean – Mode 3 (Mean – Median) can be used. The relationship holds for a moderately skewed unimodal frequency distribution. This empirical relationship **cannot** be proved theoretically. Karl Pearson has stated it after observing it to be valid for a number of data sets by actual computations.

Quantiles

The median of the data divides the series of observations/distributions into two parts containing equal observations. There are other values which also divide the series into a number of parts. These values are called as partition values or quantiles.

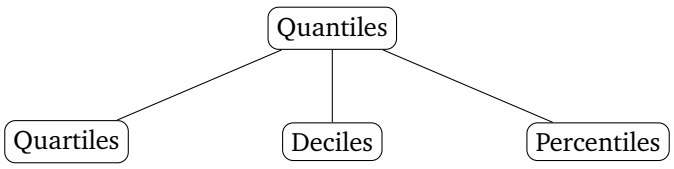

Figure 10: Basic Types of Quartiles

Quartiles

- The three values which divide the series into four parts, such that frequency of each part is 25% of the total are called quartiles.
- They are denoted by $Q_i, i = 1, 2, 3$.

- Similar to the median calculations, we simply replace $\frac{N}{2}$ by $\frac{iN}{4}$, where $i = 1, 2, 3$ denote the quartile number. Its formula is also derived from the median formula where:

$$\mathrm{Q}_i = l_1 + \frac{(l_2 - l_1)(\mathrm{iN}/4 - \text{c.f.})}{\mathrm{f}}$$

Deciles and Percentiles

The nine points which divide the series into 10 parts, such that frequency of each part is 10% of the total, are called deciles.

Whereas, percentiles are ninety nine points which divide the series into 100 parts, where frequency of each part is 1% of the total.

Thus, in particular, 6th decile will have 60% observations below it and 40% observations above it. 54th percentile will have 54% observations below it and 46% observations above it.

For raw data, $\mathrm{D}_i = i\left(\frac{n+1}{10}\right)^{\text{th}}$ observation, $\mathrm{D}_i = i\left(\frac{n+1}{100}\right)^{\text{th}}$ observation

The i$^{\text{th}}$ decile and percentile formulae for grouped frequencies respectively now are:

$$\mathrm{D}_i = l_1 + \frac{(l_2 - l_1)(\mathrm{iN}/10 - \text{c.f.})}{\mathrm{f}}$$

$$\mathrm{P}_i = l_1 + \frac{(l_2 - l_1)(\mathrm{iN}/100 - \text{c.f.})}{\mathrm{f}}$$

Comparison of Different Central Measures of Tendency

- We have studied five different measures of central tendency.
- It is **obvious** that no single measure can be the best for all situations
- The most commonly used measures are mean, median and mode.
- It is not desirable to consider any one of them to be superior or interior in all situations.
- The selection of appropriate measure of central tendency would largely depend upon the nature of the data; more specifically, on the scale of measurement for representing the data and purpose on hand.
- The data obtained on nominal scale, we can count the number of cases in each category and obtain the frequencies
- We may then be interested in knowing the class which is most typical value in the data.
- In such cases mode can be used as the appropriate measure of central tendency.

e.g. Suppose, in a genetical study, for a group of 50 family members, we want to know most common colour of eyes. Then we count the number of persons for each different colour of eye. Suppose 3 persons have light eyes, 6 persons have brown eyes, 12 with dark grey eyes and 29 persons are with black eyes. Then the more common colour of eyes (i.e. mode) for this group of people is 'black'.

- When the data is available on ordinal scale measurement i.e. the data is provided in rank order, use of median as a measure of central tendency is appropriate.
- Suppose in a group of 75 students, 10 students have failed, 15 get pass class, 20 secure second class and 30 are in first class.
- The average performance of students will be the performance of the middlemost student (arranged as per rank) i.e. the performance of 38$^{\text{th}}$ student i.e. second class; which is the median of the data.
- Median is only a point on the scale of measurement, below and above which lie exactly 50% of the data.

- Median can also be used (i) for truncated (incomplete) data, provided we know the total number of cases and their positions on the scale and (ii) when the distribution is markedly skewed
- Arithmetic Mean is the most commonly used measure of central tendency.
- It can be calculated when the data is complete and is represented on interval or ratio scale.
- It represents the centre of gravity of the data i.e. the measurement in any sample are perfectly balanced about the mean.
- In computation of simple A. M. equal importance is given to all observations in the data.
- It is preferred because of the its high reliability and its applicability to inferential statistics.
- Thus, A. M. is the more precise, reliable and stable measure of central tendency.
- Geometric mean is appropriate measure of central tendency when data is related to ratios, rates and percentages.
- It is usually used to obtain average percentage change in any characteristics.
- Geometric mean is best to be used when one wants to give more weights to small values than large values in the series
- Harmonic mean, like geometric mean, is also a measure of central tendency used in solving special types of problems.
- It is generally used for averaging speed, prices. etc.
- We have seen that each measure of central tendency has situations for its best use and also has its own limitations.
- So one has to take conscious decision about its applicability.
- Selection of appropriate measure of central tendency requires experience and insight to examine the nature of data and purpose in hand.

Unit 3: Measures of Dispersion, Skewness and Kurtosis

What makes a good measure of dispersion?

- It should be easy to calculate and simple to understand
- It should be rigidly defined
- It should be based on all observations
- It should be capable of further algebraic treatment
- It should have sampling stability
- It should not be unduly affected by extreme values
- We shall now study the various absolute and relative measures of dispersion.

There are two types of measures to look at.

Absolute Measures Range, Quartile Deviation or Semi-inter Quartile Range, Mean Deviation and Standard Deviation.

Relative Measures Coefficient of Range, Coefficient of Quartile Deviation, Coefficient of Mean Deviation and Coefficient of Variation.

Range

- If L is the largest observation in the data and S is the smallest observation, then

$$\text{Range} = L - S$$

- For a frequency distribution, range may be considered as the difference between the largest and the smallest class boundaries
- Range is crude, and the simplest measure of dispersion. It measures the scatter of observations among themselves and not about any average.
- The corresponding relative measure is

$$\text{Coefficient of Range} = \frac{L - S}{L + S}$$

- Range is a suitable measure of dispersion in case of small groups.
- In the branch of statistics known as Statistical Quality Control, range is widely used.
- It is also used to measure the changes in the prices of shares.
- Variation in daily temperatures at a certain place are measured by recording maximum temperature and minimum temperature.
- Range is also used in medical sciences to check whether blood pressure, haemoglobin count, etc. are normal.

- The main drawback of this measure is that it is based on only two extreme values, the maximum and the minimum and completely ignores all the remaining observations.x

Quartile Deviation

- We have seen earlier that range, as a measure of dispersion, is based only on two extreme values and fails to take into account the scatter of the remaining observations within the range.
- To overcome this drawback to an extent, we use another measure of dispersion called Inter-Quartile Range. It represents the range which includes the middle 50% of the distribution. Hence,

  $$\text{Inter-Quartile Range} = \mathrm{Q}_3 - \mathrm{Q}_1$$

  where Q_3 and Q_1 represent the upper and lower quartiles respectively.

  $$\text{Semi-Inter-Quartile Range} = \text{Quartile Deviation} = \frac{\mathrm{Q}_3 - \mathrm{Q}_1}{2}$$

  The SIQR or Quartile Deviation is often used.
- The corresponding relative measure is

  $$\text{Coefficient of Quartile Deviation} = \frac{\mathrm{Q}_3 - \mathrm{Q}_1}{\mathrm{Q}_3 + \mathrm{Q}_1}$$

- Q.D. is independent of extreme values. It is better representative and more reliable than range.
- Q.D. gives an idea about the distribution of middle half of the observations around the median.
- Whenever median is preferred as s measure of central tendency, quartile deviation is preferred as a measure of dispersion. However, like median, quartile deviation is also not capable of further algebraic treatment, as it does not take into consideration all the values of the distribution.
- For a symmetric distribution,

  $$\mathrm{Q}_1 = \text{Median} - \text{Q.D.}$$
  $$\mathrm{Q}_3 = \text{Median} + \text{Q.D.}$$

Mean Deviation

- Mean Deviation is defined as the arithmetic mean of the absolute deviations of the observations from any suitable constant, say A.
- Thus, if a variable X can assume values $X_1, X_2, \ldots, X_n$ and A is any arbitrary constant, then mean absolute deviation or mean deviation from A is defined as

  $$\frac{\sum_{i=1}^{n} |X_i - A|}{n}$$

- For a frequency distribution,

  $$\frac{\sum_{i=1}^{n} f_i \cdot |X_i - A|}{n}$$

- Though A is an arbitrarily selected constant, generally for statistical analysis it is taken as one of the measures of central tendency such as mean, median or mode. It is observed that M.D. from A is minimum when A is the median.
- M. D. is the simplest measure of dispersion that takes into account all the values in a given distribution. However, it has some limitations as well. First, as it takes into account the absolute deviations (i.e. does not consider signs of the deviations), it is unwieldy in mathematical operations. Secondly, it is influenced by extreme values.
- The M. D. is useful in statistical analysis of economic and social phenomena, using small samples.

Standard Deviation

- Standard Deviation (σ) is defined as the positive square root of the arithmetic mean of the square of the deviations of the observations from their arithmetic mean.

- The arithmetic mean of the squares of the deviations of the observations from their A. M. is called the variance.

$$\sigma = \sqrt{\text{Variance}}$$

For raw data,

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

For ungrouped frequency distributions,

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} f_i \cdot (x_i - \overline{x})^2}{\sum_{i=1}^{k} f_i}},$$

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i \cdot x_i}{\sum_{i=1}^{k} f_i}$$

For grouped frequency distributions,

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} f_i \cdot (x_i - \overline{x})^2}{\sum_{i=1}^{k} f_i}},$$

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i \cdot x_i}{\sum_{i=1}^{k} f_i},\ x_i\text{'s are the class marks.}$$

- The corresponding relative measure is

$$\text{Coefficient of Variance} = \frac{\sigma}{\overline{x}}$$

Properties of Standard Deviation

Effect of change of origin and scale

$$\text{If } u = \frac{x - a}{c},\ \sigma_u = \frac{\sigma_x}{c}$$

Standard Deviation for Combined Group For two populations with sizes n_1, n_2, arithmetic means $\overline{x_1}, \overline{x_2}$, and standard deviations σ_1, σ_2,

$$\overline{x} = \frac{n_1\overline{x_1} + n_2\overline{x_2}}{n_1 + n_2}$$

let $d_1 = \overline{x_1} - \overline{x}$, $d_2 = \overline{x_2} - \overline{x}$ then,

$$\sigma = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}}$$

Zero Variance If all observations in a series are equal or if the data contains only one observation, i.e. there is no variation in data, $\sigma = 0$.

Spread The spread of variable is approximately taken as $(\overline{x} \pm 3\sigma)$.

Merits

- It is rigidly defined
- It is based on all observations
- It is capable of further algebraic treatment
- It is least affected by sampling fluctuations

Demerits

- As compared to other measures, it is difficult to calculate
- It cannot be calculated for distribution with open end class-intervals
- It gives more importance (weightage) to extreme values and less importance to the values close to A. M. i.e. it is affected due to extreme observations.
- It cannot be calculated for qualitative data

Moments

There are two types of moments,

Raw Moments the deviations are taken from some constant

Central Moments the deviations are taken from A. M. of the distribution

In general, if a variable X takes values $X_1, X_2, \ldots X_n$, then the r^{th} raw moment (μ'_r) about the arbitrary origin a is calculated as

$$\mu'_r = \frac{\sum_{i=1}^{n}(x_i - a)^r}{n} \equiv \overbrace{\frac{\sum_{i=1}^{n} f_i(x_i - a)^r}{\sum_{i=1}^{n} f_i}}^{\text{For frequency distributions}}$$

Raw moments with $a = 0$ are:

$$\mu'_r = \frac{\sum_{i=1}^{n} x_i^r}{n} \equiv \overbrace{\frac{\sum_{i=1}^{n} f_i x_i^r}{\sum_{i=1}^{n} f_i}}^{\text{For frequency distributions}}$$

And the r^{th} central moment (μ_r) is

$$\mu'_r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^r}{n} \equiv \overbrace{\frac{\sum_{i=1}^{n} f_i(x_i - \overline{x})^r}{\sum_{i=1}^{n} f_i}}^{\text{For frequency distributions}}$$

Special Cases:

- $\mu'_1 = \overline{x}$
- $\mu_1 = 0$
- $\mu_2 = \sigma^2 = \mu'_2 - {\mu'_1}^2$

Skewness

The property of any deviation of frequency distribution from symmetry is known as skewness. A distribution which is not symmetrical is called a skewed distribution. The distributions representing number of accidents per day, age of individuals, number of misprints per page etc. are generally non symmetric. In case of a skewed distribution; the observations are not symmetrical placed about the average, but there are more observations on one side of the average than on the other side. Also, the mean, median and mode of the distribution do not coincide, and the quartiles are not equi-spaced.

A frequency distribution is symmetric about some point $x = a$, if the spread of frequencies on both sides of a is the same. This means that if a frequency curve of such a distribution is folded on the ordinate $x = a$, the two halves of the curve will coincide with each other.

Types of Skewness:

Symmetric

- For a symmetric distribution, the values of mean, median and mode coincide, and also the lower and upper quartiles are equi-spaced from the median.
- A symmetric distribution has zero skewness.

Positive

- If the density of observations is more for the lower values of the variable than for higher values of the variables, then the frequency curve increases rapidly to reach the maximum and further decreases slowly.
- The tail of the frequency curve is elongated towards the right hand side.
- Mean > Median > Mode, Q_3 - Q_2 > Q_2 - Q_1,
- $\mu_3 > 0$

- The distribution of income of individuals, house-rent, profits of companies etc. are usually positively skewed.

Negative

- If the density of observations is more for higher values of the variable than for lower values of the variable
- The tail of the frequency curve is elongated towards the left side.
- Mean < Median < Mode, Q_3 - Q_2 < Q_2 - Q_1,
- $\mu_3 < 0$
- The distribution of supply with respect to price, numbers of depositors with respect to savings in the bank are usually negatively skewed.

Absolute Measures of Skewness

Pearson's Measure

$$= \text{Mean} - \text{Mode}$$

Bowley's Measure

$$= (Q_3 - Q_2) - (Q_2 - Q_1))$$

Relative Measures of Skewness

Pearson's Coefficient

$$SK_p = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

If SK_p is positive, the data is positively skewed, and vice versa.

Bowley's Coefficient

$$SK_B = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

If SK_B is positive, the data is positively skewed, and vice versa. Also, $|SK_B| < 1$.

Relative Measure based on Moments

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}, \gamma_1 = \sqrt{\beta_1}$$

If $\gamma_1 = 0$, the distribution is symmetric.
If $\gamma_1 > 0$, the distribution is positively skewed.
If $\gamma_1 < 0$, the distribution is negatively skewed.

Kurtosis

Kurtosis is the other important aspect of a curve which refers to its degree of flatness or peakedness. The frequency curves of different distributions have different degree of flatness or peakedness about the mode. If the spread of the observations in a distribution is less, then there is more concentration of observations around mode and the frequency curve of such a distribution will have a taller peak as compared to the frequency curve of a distribution having more spread. In short, frequency curves can be broadly categorized as of three different types with regard to the convexity, the property which is referred to as **Kurtosis**.

The kurtosis of a bell shaped curve (of normal distribution) is taken as standard. Such as curve is neither peaked nor flat-topped and is called a 'mesokurtic' curve. A curve which is more flat-topped than normal curve is called 'platykurtic' and a curve which is more peaked than the normal curve is called as 'leptokurtic'.

Measures of Kurtosis Relative Measure Based on Moments

$$\beta_2 = \frac{\mu_4}{\mu_2^2}, \gamma_2 = \beta_2 - 3$$

If $\gamma_2 = 0, \beta_2 = 3$, the distribution is mesokurtic.
If $\gamma_1 > 0, \beta_2 > 3$, the distribution is leptokurtic.
If $\gamma_1 < 0, \beta_2 < 3$, the distribution is platykurtic.

Box-and-Whisker Plot A box and whisker display (a.k.a. boxplot) is another graphic technique used in Exploratory Data Analysis. It shows the five number summary for a univariate data set, giving a quick impression of the distribution. The five number summary is defined as a set of five landmark points of a data viz. smallest data value, lower quartile, median, upper quartile and the largest data value.

Step 1. To construct a horizontal box and whisker diagram, draw a horizontal axis that is scaled to the data. Above the axis, draw a rectangle box of convenient height with the left side (or left hinge) and right side (or right hinge) at Q_1 and Q_3 respectively.

Step 2. Draw a vertical line segment at Q_2. The length of this box is then equal to Q_3 - Q_1 i.e. the interquartile range (IQR).

Step 3. Then, two sets of limits viz. inner fence and outer fence are calculated as ($Q_1 - 1.5\times$ IQR), ($Q_3 + 1.5\times$ IQR) and ($Q_1 - 3\times$ IQR), ($Q_3 + 3\times$ IQR) respectively.

Step 4. Two lines are drawn outside the box parallel to horizontal axis, one from the midpoint of the left side of the box to the smallest value inside lower limit of inner fence and the other from the midpoint of the right side of the box to the largest value inside upper limit of the inner fence.

Step 5. These two lines are termed as the left whisker and right whisker respectively.

Sometimes, the given data is such that few observations are much larger or much smaller than most of the observations. Such observations are called outliers are usually indicated by special symbols such as * on the box plot. Such observations, if they lie between inner and outer fence, are called suspected outliers. The observations beyond the outer fence are called outliers.

The box and whisker display can show, at a glance, whether the given data set is reasonably symmetrical or skewed. A median line that is approximately in the centre of the box indicates that the data is reasonably symmetric. But a median line towards the right side of the box indicated negative skewness, while a median line towards the left of the box suggests positive skewness. Also, skewness is indicated if one whisker line is appreciably longer than the other Box-plots are useful for comparing several groups of numbers.

Part II

Descriptive Statistics - II

Syllabus

Unit 1: Theory of Attributes

1. Dichotomous classification
2. Association of attributes:
   - Yule's coefficient of association
   - Odds Ratio
   - Gamma coefficient

Unit 2: Correlation and Regression

1. Correlation Analysis
   - Scatter Diagram
   - Product moment correlation coefficient and its properties
   - Spearman's rank correlation coefficient
   - Spurious correlation
2. Regression Analysis
   - Principle of least squares
   - Fitting a straight line by method of least squares
   - Concept and use of coefficient determination r^2
   - Fitting of curves reducible to linear form by transformation
   - Fitting a quadratic curve by method of least squares

Unit 3: Index Numbers

1. Index numbers as a comparative tool
2. Stages in the construction of Price Index Numbers
3. Measures of Simple and Composite Index Numbers
   - Laspeyre's
   - Paasche's
   - Marshal-Edgeworth's
   - Drobisch and Bowley's
   - Fisher's
4. Quantity Index Numbers and Value Index Numbers
5. Time reversal test
6. Factor reversal test
7. Circular test
8. Fixed base Index Numbers
9. Chain base Index Numbers
10. Base shifting, splicing and deflating
11. Cost of Living Index Number
12. Concept of Real Income based on Wholesale Price Index Number

Unit 1: Theory of Attributes

In many situations, two qualitative characteristics or attributes are studied with the objective to know whether any kind of association exists between them.
e.g. we would like to know whether education of a student and area of residence are related or inoculation and prevention of disease are associated.

Association of Attributes

- Let A and B denote two attributes under study
- Each attribute is divided into two disjoint classes. A and α denote presence and absence of attribute A while B and β denote presence and absence of attribute B respectively
- Let N denote the total number of items under study
- Let (A) denote number of items possessing attribute A and (α) denote number of items not possessing attribute A
- Similarly, (B) and (β) denote number of items possessing and not possessing attribute B respectively

$$(A) + (\alpha) = (B) + (\beta) = N$$

- Also, let (AB) denote the number of items possessing both A and B
- Similarly (Aβ), (αB) and ($\alpha\beta$) denote number of items possessing A but not possessing B, not possessing A and possessing B and not possessing A and B both, respectively

$$(AB) + (A\beta) = (A)$$
$$(\alpha B) + (\alpha\beta) = (\alpha)$$
$$(AB) + (\alpha B) = (B)$$
$$(\alpha B) + (\alpha\beta) = (\beta)$$

The following contingency tables express these forms.

Order and Class Frequency

The order of a class depends upon the number of attributes specified. A class having one attribute is known as the class of the first order, a class having two attributes as class of the second order, and so on. The total number of observations denoted by the symbol N is called the frequency of the zero order since no attributes are specified.

A \ B	B	β	Total
A	(AB)	(Aβ)	(A)
α	(αB)	($\alpha\beta$)	(α)
Total	(B)	(β)	N

Table 6: 2×2 contingency table

Attribute	Attribute	C	γ	Total
A	B	(ABC)	(ABγ)	(AB)
	β	(AβC)	(A$\beta\gamma$)	(Aβ)
	Total	(AC)	(Aγ)	(A)
α	B	(αBC)	(αBγ)	(αB)
	β	($\alpha\beta$C)	($\alpha\beta\gamma$)	($\alpha\beta$)
	Total	(αC)	($\alpha\gamma$)	(α)
Total	B	(BC)	(Bγ)	(B)
	β	(βC)	($\beta\gamma$)	(β)
	Total	(C)	(γ)	N

Table 7: 3×3 contingency table

For a 2×2 contingency table,

Zero Order N

First Order (A), (B), (α), (β)

Second Order (AB), (Aβ), (αB), ($\alpha\beta$),

and are known as the ultimate frequencies

In a study of n attributes, the total number of class frequencies are 3^n.

Consistency of Data

- In order to find whether the given data are consistent or not we have to apply a very simple test
- The test is to find out whether any one or more of the ultimate class-frequencies is negative or not
- If none of the class-frequencies is negative we can safely conclude that the given data are consistent (i.e. The frequencies do not conflict in any way with each other)
- On the other hand, if any of the ultimate class-frequencies comes out to be negative then the given data are inconsistent
- Thus the necessary and sufficient condition for the consistency of a set of independent class-frequencies is that no ultimate class frequency is negative

Independence

Now, to understand data further we need to answer the following questions

- Is there any association between the two attributes?
- If yes, then what is the type of association?
- What is the amount/extent of this association?

Types of Association

- Two attributes A and B are said to be independent if there does not exist any association between them
- Mathematically, A and B are independent if $\frac{(AB)}{B} = \frac{(A\beta)}{\beta}$, (AB)($\alpha\beta$) = (A$\beta$)($\alpha$B), or (AB) = $\frac{(A)(B)}{N}$.
- A and B are said to be positively associated if (AB) $> \frac{AB}{N}$, and negatively associated if (AB) $< \frac{AB}{N}$.
- If A always occurs with B and never with β i.e. (AB) = (A) or (Aβ) = 0, or B always occurs with A and never with α i.e. (AB) = (B) or (αB)=0, then A and B are said to be completely associate
- On the other hand, if A never occurs with B i.e. (AB) = 0 or α never occurs with β i.e. ($\alpha\beta$) = 0, then attributes A and B are said to be completely disassociated.

Measuring Degree of Association

Yule's Coefficient (Q)

- It is the most popular method of determining not only nature of association but also the degree or extent to which the two attributes are associated.

$$\mathrm{Q} = \frac{(\mathrm{AB})(\alpha\beta) - (\mathrm{A}\beta)(\alpha\mathrm{B})}{(\mathrm{AB})(\alpha\beta) + (\mathrm{A}\beta)(\alpha\mathrm{B})}$$

- $-1 \leq \mathrm{Q} \leq 1$
- Interpretations:

  If Q $= 0$
  A and B are independent

  If Q $= 1$
  A and B are completely associated

  If Q $= -1$
  A and B are completely disassociated

  If Q $\in (-1, 0)$
  A and B are negatively associated

  If Q $\in (0, 1)$
  A and B are positively associated

Coefficient of Colligation (Y)

$$\mathrm{Y} = \frac{\sqrt{(\mathrm{AB})(\alpha\beta)} - \sqrt{(\mathrm{A}\beta)(\alpha\mathrm{B})}}{\sqrt{(\mathrm{AB})(\alpha\beta)} + \sqrt{(\mathrm{A}\beta)(\alpha\mathrm{B})}}$$

- $-1 \leq \mathrm{Y} \leq 1$
- Interpretations:

  If Y $= 0$
  A and B are independent

  If Y $= 1$
  A and B are completely associated

  If Y $= -1$
  A and B are completely disassociated

  If Y $\in (-1, 0)$
  A and B are negatively associated

  If Y $\in (0, 1)$
  A and B are positively associated

Show the relationship between Q and Y.

$$\begin{aligned}\mathrm{Y} &= \frac{\sqrt{(\mathrm{AB})(\alpha\beta)} - \sqrt{(\mathrm{A}\beta)(\alpha\mathrm{B})}}{\sqrt{(\mathrm{AB})(\alpha\beta)} + \sqrt{(\mathrm{A}\beta)(\alpha\mathrm{B})}} \\ &= \frac{1 - \frac{\sqrt{(\mathrm{A}\beta)(\alpha\mathrm{B})}}{\sqrt{(\mathrm{AB})(\alpha\beta)}}}{1 + \frac{\sqrt{(\mathrm{A}\beta)(\alpha\mathrm{B})}}{\sqrt{(\mathrm{AB})(\alpha\beta)}}}\end{aligned}$$

Let $k = \frac{\sqrt{(\mathrm{A}\beta)(\alpha\mathrm{B})}}{\sqrt{(\mathrm{AB})(\alpha\beta)}}$

$$
\begin{aligned}
&= \frac{1-\sqrt{k}}{1+\sqrt{k}} \\
\mathrm{Y}^2 &= \frac{1-2\sqrt{k}+k}{(1+\sqrt{k})^2} \\
1+\mathrm{Y}^2 &= \frac{2(1+k)}{(1+\sqrt{k})^2} \\
&= \frac{2(1+k)}{(1+\sqrt{k})^2} \\
\frac{2\mathrm{Y}}{1+\mathrm{Y}^2} &= 2\,\frac{1-\sqrt{k}}{1+\sqrt{k}}\,\frac{(1+\sqrt{k})^2}{2(1+k)} \\
&= \frac{1-k}{1+k} \\
&= \frac{1-\frac{(\mathrm{A}\beta)(\alpha\mathrm{B})}{(\mathrm{AB})(\alpha\beta)}}{1+\frac{(\mathrm{A}\beta)(\alpha\mathrm{B})}{(\mathrm{AB})(\alpha\beta)}} \\
&= \frac{(\mathrm{AB})(\alpha\beta)-(\mathrm{A}\beta)(\alpha\mathrm{B})}{(\mathrm{AB})(\alpha\beta)+(\mathrm{A}\beta)(\alpha\mathrm{B})} \\
\frac{2\mathrm{Y}}{1+\mathrm{Y}^2} &= \mathrm{Q}
\end{aligned}
$$

Odds Ratio

- An odds ratio (OR) is a measure of association between an exposure and an outcome
- The OR represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure

$$\text{Odds} = \frac{\mathrm{P}_{\text{event}}}{1-\mathrm{P}_{\text{event}}}$$

- The odds ratio shows the strength of the association between the predictor variable (exposure) and the outcome variable (response).
- If the odds ratio is 1, then there is no association between the predictor variable and the outcome.
- If the odds ratio is greater than 1, then exposure is associated with higher odds of outcome
- If the odds ratio is less than 1, then exposure is associated with lower odds of outcome

Goodman and Kruskal's Gamma

Gamma is $\overbrace{\text{a measure whose value will be the same when either attribute is considered independent or dependent.}}^{\text{a symmetric measure of association}}$

A gamma coefficient tells us how closely two pairs of data points "match", as well as the strength of association.

In Table 8, concordant pairs (N_c) are calculated from $N_c = ad$, and discordant pairs $N_d = bc$. The Gamma coefficient is then defined as:

$$\gamma = \frac{N_c - N_d}{N_c + N_d} = \frac{ad-bc}{ad+bc}$$

		Independent	
	B / A	B_1	B_2
Dependent	A_1	a	b
	A_2	c	d

Table 8: Sample 2×2 contingency table

	Independent		
B / A	B_1	B_2	B_3
A_1	a	b	c
A_2	d	e	f
A_3	g	h	i

Table 9: Sample 3×3 contingency table

In Table 9,

$$N_c = \mathrm{a(e+f+h+i)+b(f+i)+d(h+i)+ei}$$
$$N_d = \mathrm{c(d+e+g+h)+b(d+g)+f(g+h)+eg}$$
$$\gamma = \frac{N_c - N_d}{N_c + N_d}$$

Interpretations:

If $\gamma > 0$

Positive Association

If $\gamma < 0$

Negative Association

If $\gamma \in [0, 0.25)$

No relationship

If $\gamma \in [0, 25, 0.50)$

Weak relationship

If $\gamma \in [0.50, 0.75)$

Moderate relationship

If $\gamma \in [0.75, 1)$

Strong relationship

Pearson's χ^2 Statistic

If we are not interested in the strength of association but rather in finding out whether there is an association at all, one can use the χ^2 independence test. Pearson's χ^2 statistic is used for measuring the association between variables in a contingency table and plays an important role in the construction of statistical tests. It is also symmetric.

For the contingency table shown in Table 10, χ^2 is defined as

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(n_{ij} - \tilde{n}_{ij})^2}{\tilde{n}_{ij}} = \frac{(n_{ij} - \frac{n_{i+}n_{+j}}{n})^2}{\frac{n_{i+}n_{+j}}{n}}$$

		y					
		y_1	$\dots$	y_j	$\dots$	y_l	Total
x	x_1	n_{11}	$\dots$	n_{1j}	$\dots$	n_{1l}	n_{1+}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	x_i	n_{i1}	$\dots$	n_{ij}	$\dots$	n_{il}	n_{i+}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	x_k	n_{k1}	$\dots$	n_{kj}	$\dots$	n_{kl}	n_{k+}
	Total	n_{+1}	$\dots$	n_{+j}	$\dots$	n_{+l}	n

Table 10: $k \times l$ contingency table

where n_{ij} are the observed frequencies, and $\tilde{n}_{ij}$ are the expected frequencies.

Note:

$$0 \leq \chi^2 \leq n(\min\{k, l\} - 1)$$

The closer the value of χ^2 is to $n(\min\{k, l\} - 1)$, the stronger the association.

Cramer's V Normalising χ^2 to lie between 0 and 1, we get

$$\mathrm{V} = \sqrt{\frac{\chi^2}{n(\min\{k, l\} - 1)}}$$

Note:

$$0 \leq \mathrm{V} \leq 1$$

The closer the value of V is to 1, the stronger the association.

Contingency Coefficient For determining the degree of association between two attributes on the whole, the coefficient of contingency (C) as given by Pearson may be used.

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Alternatively,

$$C_{\mathrm{corr}} = \frac{C}{C_{\max}} = \sqrt{\frac{\chi^2}{\chi^2 + n}} \cdot \sqrt{\frac{\min\{k, l\}}{\min\{k, l\} - 1}}$$

where

$$C_{\max} = \sqrt{\frac{\min\{k, l\} - 1}{\min\{k, l\}}}$$

Unit 2: Correlation and Regression

Correlation

- If the movement in one variable is associated with the movement in the other variable, the variables are said to be correlated while the movement in one is not associated with other then there is no correlation between the variables
- Correlation analysis is the statistical tool to identify and measure the strength of relationship between variable
- The study of correlation between two variables is called simple correlation, while that of more than two variables is called multiple correlation
- The first step in determining whether there is a relationship between two variables is to collect the values of the variables at different places or time points, or for different individuals

e.g. price and demand of certain commodity may be observed in different cities, weight and calorie intake are noted down for a few patients etc.

- These are paired observations on two variables and are denoted by (x, y)
- The graph of observed values of two variables is used to identify the relationship between two variables
- Such a graph is called a scatter diagram
- Usually, one variable depends to some degree on the other, (weight depends on calorie intake) then the dependent variable is taken on the vertical (Y) axis and the other variable is taken on horizontal (X) axis
- The pattern of points plotted on the scatter diagram indicates whether the variables are related
- The wider scattering indicates that there is low degree of correlation between two variables
- The pattern of points also indicates whether a straight line relationship exists between the two variables or the relationship between the variables can take the form of a curve (curvilinear relationship)

Karl Pearson's Product Moment Correlation Coefficient

Karl Pearson's correlation coefficient (r) quantifies the strength of the linear relationship between two variables measured on interval or ratio scale.

$$r = \frac{\text{cov}(x, y)}{\sigma_x \, \sigma_y}$$

Interpretations:

$r = 0$ There is no correlation

$r = 1$ There is perfect direct correlation

$r = -1$ There is perfect inverse correlation

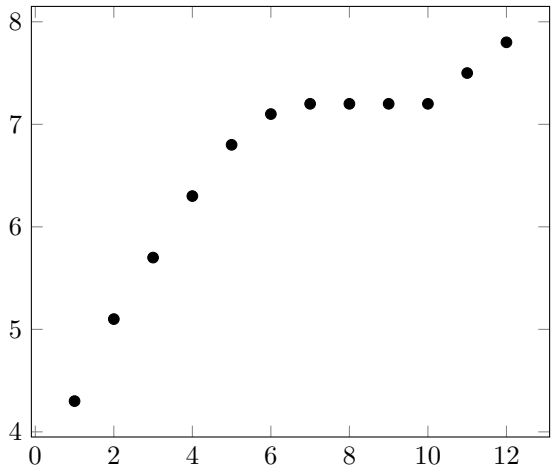

Figure 11: Scatter Plot

Properties:

- $-1 \leq r \leq 1$
- r is not affected by change in origin and scale i.e. if $u = \dfrac{x-a}{c}$, $v = \dfrac{y-b}{d}$, then $r_{u,v} = r_{x,y}$
- The correlation coefficient is a pure number. It is independent of unit of measurement of the variables

Spearman's Rank Correlation

Spearman's Rank Coefficient is a special case of Pearson's P.M.C.C, for ranking systems. It is defined as

$$r = 1 - \frac{6\sum_{i=1}^{n} d^2}{n(n^2-1)}$$

where $d = R_1 - R_2$, R_1 is the rank in the first set, R_2 is the rank in the second set, n is the number of observations in each set.

- While assigning ranks, if some individuals have the same merit, then they should be given the same rank.
- This common rank is the mean of the ranks that they would have been given if they were of different merits.
- In case of such tied ranks, $\sum d^2$ need to be corrected by adding correction factor to $\sum d^2$ for every repeated rank
- The correction factor is $\text{c.f.} = \frac{m(m^2-1)}{12}$, where m is the number of times a rank is repeated, then

$$r = 1 - \frac{6\sum_{i=1}^{n}(d^2 + \text{c.f.})}{n(n^2-1)}$$

Kendall's Tau

$$\tau = \frac{C-D}{C+D}$$

For continuous variables,

$$\tau = \frac{C-D}{{}^{n}C_2}$$

Properties

- The denominator is the total no. of pair combinations, so that coefficient must be in the range $-1 \leq \tau \leq 1$
- If the agreement between the two rankings is perfect (i.e. the two rankings are the same), the coefficient has the value 1
- If the disagreement between the two rankings is perfect (i.e. one ranking is the reverse of the other) , the coefficient has the value -1
- If X and Y are independent then we would expect the coefficient to be approximately zero

Bivariate Frequency Distributions

- Our data collection exercise would be getting paired observations $(x_1, y_1)(x_2, y_2) \ldots (x_n, y_n)$ on variables X and Y.
- The frequency distributions obtained as a result of this cross classification gives rise to a bivariate frequency distribution
- The bivariate frequency table is a two-way table

Regression

- Regression analysis develops an equation or mathematical formula between the related variables
- If the equation is established between the variables, it helps in predicting, controlling and planning the activities

e.g. if the quantity demanded depends on price of the commodity, knowing the price amount of quantity demanded can be predicted

- If expiry date of a medicine is correlated to temperature of the place where it is stored then expiry date can be controlled by temperature
- If the duration of exercise is related to cholesterol level, the patient can plan the exercise schedule to maintain a cholesterol level
- Regression analysis develops an equation which represents the observed values of two variables plotted on the scatter diagram
- Hence the graph of the equation should be as close as possible to the points in the scatter diagram
- Such an equation is called the regression equation
- It is used to predict the value of one variable given the value of the other variable
- If the variable being predicted (dependent variable) is y and the variable being used to predict the value of the dependent variable (independent variable) is x, then regression equation is called regression equation y on x and is used to estimate value of y when value of x is known
- The scatter diagram is drawn by taking the dependent variable on the vertical (Y) axis and the other variable on horizontal (X) axis
- It may reveal either a linear or curvilinear relationship between two variables
- If the scatter diagram indicates linear relationship then a linear equation between two variables is established and is called a simple linear regression equation

Simple Linear Regression

- The form of linear regression equation of y on x is

$$y = a + bx$$

  where x and y are the variables, a and b are constants

- The linear equation (straight line equation) should be as close as possible to the points in scatter diagram i.e. a line of good fit is desired
- Obtaining a line of good fit is to determine the values of the constants (a and b) in the linear equation such that the estimates of y based on the linear regression equation $\hat{y} = a+bx$ is close to the observed values of y i.e. the error $e = y - \hat{y} = y - a - bx$ should be negligible
- The errors are the distances or deviations of the observed values from the estimated values obtained from the regression line measured along the y axis
- By minimising sum of squares of errors, we get the following system of equations

$$\sum y = na + b\sum x$$
$$\sum xy = a\sum x + b\sum x^2$$

  Solving these, we get the values for a and b for which SSE is minimised.

- In this, $b = \frac{\text{cov}(x,y)}{\sigma_x^2}$ and is known as the regression coefficient of y on x (b_{yx})
- Hence, $a = \overline{y} - \overline{x} \cdot b_{yx}$

Coefficient of Determination

- The coefficient of determination is r^2, where r is Karl Pearson's correlation coefficient
- Therefore coefficient of determination $0 \leq r^2 \leq 1$
- It is calculated to judge how well the estimated regression equation fits the data
- The coefficient of determination measures goodness of fit for the estimated regression equation
- It is proportionate variation in y due to x
- Hence, higher value of coefficient of determination signifies the regression equation is a good fit to the given data

Properties

- If x is assumed to be independent and y depends on x then regression equation is called regression equation y on x and is given by

$$y - \overline{y} = b_{yx}(x - \overline{x})$$

  where b_{yx} is the regression coefficient of y on x, and is defined as

$$b_{yx} = \frac{\text{cov}(x,y)}{\sigma_x^2}$$

- If y is independent and x depends on y, then regression equation is called regression equation x on y and is given by

$$x - \overline{x} = b_{xy}(y - \overline{y})$$

  where b_{xy} is the regression coefficient of y on x, and is defined as

$$b_{yx} = \frac{\text{cov}(x,y)}{\sigma_y^2}$$

- The point $(\overline{x}, \overline{y})$ lies on both regression equations. It is the point of intersection of two regression equations
- If there is perfect correlation between x and y i.e. r = 1 or -1 then the two regression equations are same or regression lines coincide. If there is no correlation between x and y i.e. r = 0 then the two regression lines are $x = \overline{x}$ and $y = \overline{y}$ i.e. the two regression lines are perpendicular to each other.
- if $u = \frac{x-a}{c}$, $v = \frac{y-b}{d}$, then $b_{yx} = \frac{d}{c} b_{vu}$

Curve Fitting

The observed values of the two variables need not always show linear relationship between two variables. In such cases, we must fit curves to the data.

Types of common curves, with their equations and system of equations to minimise SSE

Quadratic

$$y = a + bx + cx^2$$

$$\begin{aligned} \sum y &= na + b\sum x + c\sum x^2 \\ \sum xy &= a\sum x + b\sum x^2 + c\sum x^3 \\ \sum x^2 y &= a\sum x^2 + b\sum x^3 + c\sum x^4 \end{aligned}$$

Power

$$y = ax^b$$

Let $z = \log y$, $w = \log x$, $A = \log a$

$$z = A + bw$$

$$\begin{aligned} \sum z &= nA + b\sum w \\ \sum wz &= A\sum w + b\sum w^2 \end{aligned}$$

Exponential

$$y = ab^x$$

Let $z = \log y$, $A = \log a$, $B = \log b$

$$z = A + Bx$$

$$\begin{aligned} \sum z &= nA + B\sum x \\ \sum xz &= A\sum x + B\sum x^2 \end{aligned}$$

Logarithmic

$$y = a + b\log x$$

Let $w = \log x$

$$y = a + bw$$

$$\begin{aligned} \sum y &= na + b\sum w \\ \sum wy &= a\sum w + b\sum w^2 \end{aligned}$$

Unit 3: Index Numbers

Index numbers as a comparative tool

- Index numbers are the indicators which reflect changes in economic variables of interest over a specified period of time (prices of different commodities, industrial production, imports and exports, cost of living, sales and profits etc.)
- These indicators are required by governments, corporate bodies, international organizations etc. for policy interventions and executive decisions.
- Index numbers often measure the pressure of economic behaviours and for this reason many a times they are called economic barometers

Definition Index numbers are statistical devices designed to measure the relative change in the level of phenomenon with respect to time, geographical location, or other characteristics such as income, profession etc.

- The variable may refer to:
  i. The price of a particular commodity
  ii. The volume of quantity of a particular commodity or
  iii. The value of a commodity(price x quantity)
- Alternative Definitions:

  Wheldon/Marshal Edgeworth An index number is a device which shows by its variation the changes in a magnitude which is not capable of actual measurement in itself or of direct valuation in practice

  Lawrence Kaplan An index number is a statistical measure of fluctuations in a variable in the form of a series, and using a base period for making comparisons
- Index numbers provide a common barometer for the comparison of price, quantity and changes in value of commodities over time or place.
- They indicate the relative (mostly percentage) changes of these parameters over the time.
- An index number is constructed by expressing the price, quantity or value of an item in a current period (month, year or place) as a ratio of its price, quantity or value in the reference or base period.

Stages in Constructing Index Numbers

1. Objective
2. Selection of commodities
3. Selection of the Base Period
4. Selection of averages to be used
5. Selection of Weights
6. Data Collection
7. Availability and comparability of data

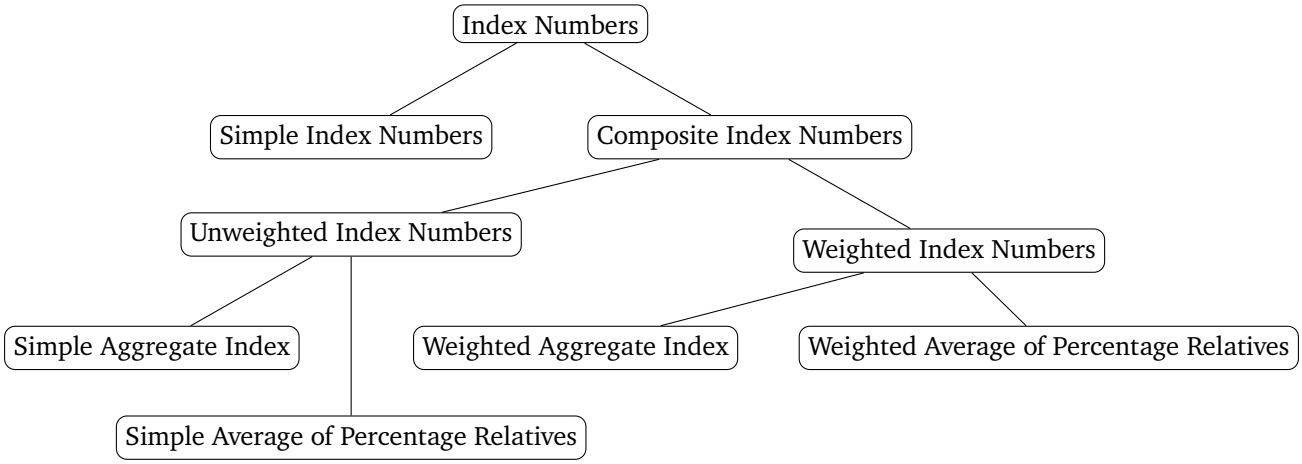

Figure 12: Simple and Composite Index Numbers

Index Numbers

Notations:

- P_0: Price in base time period
- P_t: Price in time period t
- Q_0: Quantity in base time period
- Q_t: Quantity in time period t
- V_0: Value in base time period
- V_t: Value in time period t

- Simple Index Numbers
  1. Price Relative in period t
  $$I_{t0} = \frac{P_t}{P_0} \times 100$$
  2. Quantity Relative in period t
  $$I_{t0} = \frac{Q_t}{Q_0} \times 100$$
  3. Value Relative in period t
  $$I_{t0} = \frac{V_t}{V_0} \times 100$$

- Composite Index Numbers
  1. Simple Aggregate Price Index
  $$I_{t0} = \frac{\sum P_{it}}{\sum P_{i0}} \times 100$$
  2. Simple Average of Price Values (Arithmetic Mean)
  $$I_{t0} = \frac{\sum [P_{it}/P_{i0} \times 100]}{n}$$
  3. Simple Average of Price Values (Geometric Mean)
  $$I_{t0} = \text{antilog} \left\{ \frac{\sum \log [P_{it}/P_{i0} \times 100]}{n} \right\}$$

4. Weighted Aggregate Price Index

$$I_{t0} = \frac{\sum W_i P_{it}}{\sum W_i P_{i0}} \times 100$$

5. Weighted Average of Price Values (Arithmetic Mean)

$$I_{t0} = \frac{\sum W_i [P_{it}/P_{i0} \times 100]}{\sum W_i}$$

6. Weighted Average of Price Values (Geometric Mean)

$$I_{t0} = \text{antilog} \left\{ \frac{\sum W_i \log [P_{it}/P_{i0} \times 100]}{\sum W_i} \right\}$$

7. Weighted Aggregate Value Index

$$I_{t0} = \frac{\sum P_{it} Q_{it}}{\sum P_{i0} Q_{i0}} \times 100$$

8. Weighted Aggregate Price Index (Using Quantities as Weights)

$$I_{t0} = \frac{\sum Q_i P_{it}}{\sum Q_i P_{i0}} \times 100$$

9. Weighted Aggregate Quantity Index (Using Prices as Weights)

$$I_{t0} = \frac{\sum P_i Q_{it}}{\sum P_i Q_{i0}} \times 100$$

10. Laspeyre's Price Index

$${I_{t0}}^{\mathrm{L}} = \frac{\sum Q_{i0} P_{it}}{\sum Q_{i0} P_{i0}} \times 100$$

11. Laspeyre's Quantity Index

$${I_{t0}}^{\mathrm{L}} = \frac{\sum P_{i0} Q_{it}}{\sum P_{i0} Q_{i0}} \times 100$$

12. Paasche's Price Index

$${I_{t0}}^{\mathrm{P}} = \frac{\sum Q_{it} P_{it}}{\sum Q_{it} P_{i0}} \times 100$$

13. Paasche's Quantity Index

$${I_{t0}}^{\mathrm{P}} = \frac{\sum P_{it} Q_{it}}{\sum P_{it} Q_{i0}} \times 100$$

14. Marshall-Edgeworth's Price Index

$${I_{t0}}^{\mathrm{ME}} = \frac{\sum Q_{i0} P_{it} + \sum Q_{it} P_{it}}{\sum Q_{i0} P_{i0} + \sum Q_{it} P_{i0}} \times 100$$

15. Marshall-Edgeworth's Quantity Index

$${I_{t0}}^{\mathrm{ME}} = \frac{\sum P_{i0} Q_{it} + \sum P_{it} Q_{it}}{\sum P_{i0} Q_{i0} + \sum P_{it} Q_{i0}} \times 100$$

16. Dorbish-Bowley's Price Index

$${I_{t0}}^{\mathrm{DB}} = \frac{1}{2} \left[\frac{\sum Q_{i0} P_{it}}{\sum Q_{i0} P_{i0}} + \frac{\sum Q_{it} P_{it}}{\sum Q_{it} P_{i0}} \right] \times 100$$

17. Dorbish-Bowley's Quantity Index

$${I_{t0}}^{\mathrm{DB}} = \frac{1}{2} \left[\frac{\sum P_{i0} Q_{it}}{\sum P_{i0} Q_{i0}} + \frac{\sum P_{it} Q_{it}}{\sum P_{it} Q_{i0}} \right] \times 100$$

18. Fisher's Price Index

$${I_{t0}}^{\mathrm{F}} = \left[\frac{\sum Q_{i0} P_{it}}{\sum Q_{i0} P_{i0}} \cdot \frac{\sum Q_{it} P_{it}}{\sum Q_{it} P_{i0}} \right]^{\frac{1}{2}} \times 100$$

19. Fisher's Quantity Index

$${I_{t0}}^{\mathrm{F}} = \left[\frac{\sum P_{i0} Q_{it}}{\sum P_{i0} Q_{i0}} \cdot \frac{\sum P_{it} Q_{it}}{\sum P_{it} Q_{i0}} \right]^{\frac{1}{2}} \times 100$$

Criteria of a Good Index Number

- Time Reversal Test

$$\frac{I_{t0}}{100} \times \frac{I_{0t}}{100} = 1$$

- Factor Reversal Test

$$\frac{\text{Price Index Number}}{100} \times \frac{\text{Quantity Index Number}}{100} = \frac{\text{Value Index Number}}{100}$$

- Circular Test

$$\overbrace{\frac{I_{01}}{100} \times \frac{I_{12}}{100} \times \ldots}^{n \text{ times}} \times \frac{I_{n0}}{100} = 1$$

Chain Index Numbers

Fixed Base Index Number Calculated with respect to a fixed base.

Chain Base Index Number Calculated with respect to the index number of the previous year.

Base Shifting

$$\text{Fixed Index Number in period } t = \frac{\text{Chain Index Number in period } t \times \text{Fixed Index Number in period } t-1}{100}$$

$$\text{Chain Index Number in period } t = \frac{\text{Fixed Index Number in period } t}{\text{Fixed Index Number in period } t-1} \times 100$$

To change to a different fixed base *new* from *old*,

$$I_{t,\text{ new}} = \frac{I_{t,\text{ old}}}{I_{\text{new, old}}}$$

Splicing A technique of linking two or more index number series with the same items and an overlapping year but with different base periods in order to form a single continuous index number series.

Forward Slicing

$$\text{Spliced in } t = \frac{I_{t,\text{old}}}{I_{\text{new, old}}}$$

Backward Slicing

$$\text{Spliced in } t = \frac{I_{t,\text{ new}} \times I_{\text{new, old}}}{100}$$

Cost of Living Index Number

1. Aggregate Expenditure Method

$$I_{t0}{}^{\text{L}} = \frac{\sum Q_{i0}P_{it}}{\sum Q_{i0}P_{i0}} \times 100$$

2. Family Budget Method

$$I_{t0} = \frac{\sum W_i[P_{it}/P_{i0} \times 100]}{\sum W_i}$$

where $W_i = P_{i0}Q_{i0}$

Deflating Index Numbers Deflating means 'making allowance for the effect of changing price levels'. The increase in the prices of consumer goods for a class of people over a period of years means reduction in the purchasing power for the class. For example, the increase in the price of a particular commodity from INR x in base year a to INR $2x$ in the year b implies that in year b a person can buy only half the amount of the commodity with INR x which he was spending on in year a. Thus, the purchasing power of a rupee is only 50 paise in b as compared to a.

Deflating income is done by:

$$\text{Real Income in period } t = \frac{\text{Money Wage in period } t}{\text{Price Index in period } t} \times 100$$

Wholesale Price Index Number Measures percentage change in prices of goods before the retail level i.e. goods that are sold in bulk and traded between entities and businesses instead of consumers. In some countries, it is preferred over consumer price index (CPI) to determine real income due to ease, and lesser delay, of data collection.

Unit 4: Vital Statistics

Vital statistics is accumulated data gathered on live births, deaths, migration, foetal deaths, marriages and divorces. The most common way of collecting information on these events is through civil registration, an administrative system used by governments to record vital events which occur in their populations.

Uses

Vital and Health

- To measure the state of health of a community and identify it's health problems and needs
- To compare the health status of one country with another, or to compare the present health status with that of the past
- For planning and health administration
- For evaluating the progress, success or failure of health programmes and services or operations
- For research into community health programmes

Health

- How many people suffer from a particular disease, how often or for how long
- What demands do these diseases place on medical and public health resources, what financial loss do they cause
- To what extent are people prevented by these diseases from carrying out their normal activities
- To what extent are diseases concentrating in particular groups of population according to the age, sex, occupation, place of residence etc
- How far do factors vary from time to time
- The effect of medical care and health services on the control of disease incidence

Vital

- To evaluate the impact of various national health programmes
- To plan for better future measures of disease control
- To elucidate the hereditary nature of disease
- To plan and evaluate economic and social development
- To determine the health status of an individual
- To compare the health status of one nation with others
- It is a primary tool in research activities

General

- To describe the level of community health
- To diagnose community illness
- To discover procedures, definitions, classification and techniques such as recording , systemic sampling using the skills
- To determine the priorities for health programmes
- To find clues for administrative actions

- To promote health legislation
- To create administrative standards of health activities
- To know the needs which have been met and not met
- To disseminate reliable information on the health situation and programmes
- To demand public support for health work
- To determine success or failure of specific health programme
- To quantify health problems, medical and healthcare needs
- To compare local, national and international health status of the people
- To measure the effectiveness and efficiency of accomplishing the objectives of health services
- To assess the attitudes and degree of satisfaction of the beneficiaries with the beach system
- To conduct research on particular problems of health and disease
- To measure the helth status of the people

Sources

Vital Health

Census The total process of collecting, compiling and publishing demographic, economic ad social data pertaining to a specific time or times to all persons in a country delimited territory

Registration of Vital Events It is the legal registration, statistical recoding and reporting of the occurrence of statistics and the collection, compilation, presentation, analysis and distribution of statistics pertaining to vital events i.e. live births, death, foetal deaths, marriages, divorces, adoptions, legitimation, recognition, annulment and legal separations

Sample Registration System (SRS) It is used to provide reliable estimate of birth and death rates at a national and state level. It is also a dual record system, consisting of continuous enumeration of births and deaths by an enumerator and an independent survey every 6 months by an investigator supervisor.

Notification of Disease Notification provides valuable information about fluctuations in disease frequency. It also provides early warning about new occurences or outbreaks of disease.

Hospital Records The hospital record provides information about age, sex, diagnosis, time interval between occurrence and hospital admission and distribution of patients according to different social and biological characteristics.

Disease Register Morbidity register enlists not only certain diseases and conditions but also provides information about duration of illness, case fatality and survival. These registers allow follow up patients and provide a continuous account of the frequency of disease in the community.

Record Linkage Medical record linkage implies the assemble and maintenance for each individual in a population of a file of the more important records relating to his/her health. The events commonly recorded are birth, marriage, death, hospital admission and discharge.

Epidemiological Surveillance It is a system used to report on the occurrence of new cases and on efforts to control diseases.

Other Health Service Records It includes outpatient department, primary healthcare centres. Subcenter, polyclinics, private practitioners, MCH centres, school health records, diabetic and hypertensive clinics etc.

Environmental Health Data It is helpful in the identification and qualification of causative factors of disease. Collection of environmental data plays an essential role to ascertain major problems for the future.

Manpower statistics It is the information about the number of physicians, dentists, pharmacists, veterinarians, hospital medical technicians. Their records are maintained by the state medical/dental/nursing council and the directorate of medical education.

Vital

Health Surveys Surveys are carried out for epidemiological diseases in the field by the practitioners to find the incidence and prevalence of the disease situation in a community. They also study the merits of different methods adopted to control disease.

Records of Hospitals and Health Centres Records are maintained as a routine in registers over a long period of time for various purposes such as vital statistics like births, marriages, deaths in hospital. Data for these records are used for demography and public health practice. The data collected at local, state and national level may be analysed to assess changes in the disease situation in the community.

Health

Medical and Health Centres They provide services to the community such as hospitals, clinics, dispensaries, maternal child health centres, diagnostic laboratories, mass immunisation centres, medical care programmes. They provide the statistics of admission to hospitals and other medical institutions.

Surveys They are conducted in response to the need for more detailed information con geographical basis such as nutritional surveys and epidemiological investigation in various diseases.

Physicians Care It records while the police reports the data on accidents affecting the health of the population.

Journals They have periodic publications about vital events.

Various Rates

Crude Death Rate The crude death rate is calculated as the number of deaths in a given period divided by the population exposed to risk of death in that period.

Specific Death Rate Cause specific death rate is the number of deaths from a specified cause per 100,000 person-years at risk.

Standardised Death Rate The standardised death rate, abbreviated as SDR, is the death rate of a population adjusted to a standard age distribution. It is calculated as a weighted average of the age-specific death rates of a given population; the weights are the age distribution of that population.

Crude Birth Rate The crude birth rate is generally computed as a ratio. The numerator is the number of live births observed in a population during a reference period and the denominator is the number of person-years lived by the population during the same period. It is expressed as births per 1,000 population.

Age Specific Fertility Rate Age-specific fertility rate refers to the number of births to females in a particular age category in a particular year compared to the number of females in that age category. Female refers to a person whose sex is female.

General Fertility Rate The general fertility rate (GFR) is the average number of children currently being born to women of reproductive age in the period, typically 1-36 months preceding the survey, expressed per 1,000 women age 15-44.

Total Fertility Rate It is expressed as births per woman. The total fertility rate is the sum of the age-specific fertility rates for all women multiplied by five.

Gross Reproduction Rate The gross reproduction rate is the average number of daughters a woman would have if she survived all of her childbearing years, which is roughly to the age of 45.

Net Reproduction Rate In population ecology and demography, the net reproduction rate, R0, is the average number of offspring (often specifically daughters) that would be born to a female if she passed through her lifetime conforming to the age-specific fertility and mortality rates of a given year.