Master of Science in Mathematics
(M.Sc. Mathematics)

Biostatistics
(DMSMVA201T24)

Self-Learning Material
(SEM II)

Jaipur National University
Centre for Distance and Online Education

Established by Government of Rajasthan
Approved by UGC under Sec 2(f) of UGC ACT 1956
&
NAAC A+ Accredited

TABLE OF CONTENTS

EXPERT COMMITTEE

1. **Dr. Vikas Gupta**
Dean
Department of Mathematics
LNMIIT, Jaipur

2. Dr. Nawal Kishor Jangid
Department of Mathematics
SKIT, Jaipur

COURSE COORDINATOR

Mr. Nitin Chouhan
Dept. of Basic Sciences
JNU, Jaipur

UNIT PREPARATION

Unit Writers	Assisting & Proofreading	Unit Editor
Mr. Nitin Chouhan Dept. of Basic Sciences JNU, Jaipur Unit 1 - 3	Prof. Hoshiyar Singh Dept. of Basic Sciences JNU, Jaipur	Dr.Yogesh Khandelwal Dept. of Basic Sciences JNU, Jaipur
Dr. Yogesh Khandelwal Dept. of Basic Sciences JNU, Jaipur Unit 4 - 8		
Mr. Mohammed Asif Dept. of Basic Sciences JNU, Jaipur Unit 9 - 10		

Secretarial Assistance:

Mr. Mukesh Sharma

COURSE INTRODUCTION

Biostatistics is a branch of statistics focused on the application of statistical methods to biological, medical, and health-related research. It involves the design of experiments, data collection, and the analysis and interpretation of data to make inferences about biological phenomena.

The course provides students with a comprehensive understanding to use techniques such as hypothesis testing, regression analysis, and survival analysis to make sense of complex data and draw valid conclusions. Their work supports evidence-based decisions in clinical trials, epidemiological studies, and health policy. By providing insights into the effectiveness of treatments, risk factors, and disease patterns, biostatistics plays a crucial role in advancing scientific knowledge and improving health outcomes.

The course is of two credits and divided into 10 units. Each Unit is divided into sections and subsections in each unit. Each unit starts with a statement of objectives that outlines the goals we hope you will accomplish.

Course Outcomes:

At the completion of the course, a student will be able to:

1. Recall the Mean, Median, Mode and Measures of Location Quartiles.
2. Explain the Range Deviation, Quartile Deviation etc., MeanDeviation and Variance.
3. Apply the Mean, Median, Mode methods.
4. Analyze the Measures of Central Tendency and Variation forQualitative Variables.
5. Evaluate integrals of vector valued function over curves,surfaces and domains in two and three-dimensional space.
6. Create Graphical Representation by types of data for univariate and bivariate presentation.

Acknowledgements:

The content we have utilized is solely educational in nature. The copyright proprietors of the materials reproduced in this book have been tracked down as much as possible. The editors apologize for any violation that may have happened, and they will be happy to rectify any such material in later versions of this book.

UNIT - 1

Introduction to Graphical Representation

Learning Objectives

- Understand the data visualization
- Understand the Collection of secondary data
- Understand Selection of appropriate methods of data collection

Structure

1.1 Overview of data visualization

This is one of the most common methods of data collection. It is used by common individuals, researchers, organizations and government institutions. This approach involves giving the respondents a questionnaire to complete, following which the questionnaire is collected. A questionnaire is made up of several questions in a predetermined order. The respondents have to answer the questions by themselves and without any external help. The questionnaires can also be mailed to the respondents. Advantages: In case of data collection from large area this method is cost effective. This method is unbiased because there is no interviewer to affect the answers whatsoever. This method gives adequate time to respondents to answer. Limitations: In case of the mailed questionnaires there is low rate of return of the duly filled in questionnaires and sometimes very few questions are answered properly. This method requires a certain level of education and cooperation among the respondents. In case of the mailed questionnaires we cannot know for sure that the answers given by respondents are actually given by them or someone else filled their questionnaires. Precautions: We should carry out a pilot survey to test the questionnaires prior to utilizing the questionnaire approach. The pilot survey serves as the primary survey's practice run. Through the pilot survey we can understand the flaws in our technique and also can modify the questionnaire according to our respondents.

1.2 Collection of secondary data

As we have seen earlier that collection of primary data is a tedious task and it involves a lot of money, time and manpower. Hence to save all the trouble sometimes we collect data which are already available in published and unpublished form, called as Secondary data. These secondary data are collected by some organization or individual in the past, so we just go to that source and collect that data to avoid the problems associated with the collection of original data. We can collect published data from different reports of foreign, central, state and local governments, international bodies, various journals, books, magazines, newspapers, historical documents and other sources of published information. We can collect published data from diaries, letters, unpublished biographies and autobiographies, unpublished research work etc. We should be very careful in applying the secondary data to our research because that data have not been collected according to our objectives. Secondary data might not be suitable or adequate in the context of our research. Before using the data we should check the origin point of data, source of data, the

collection methodology of data, the time of data collection, bias in the compiler and the level of accuracy. Only after getting ensured about all these points we can use the data.

1.3 Selection of appropriate methods of data collection

Data collection is a crucial step in statistical analysis, as the quality and reliability of the data gathered directly impact the validity of the statistical findings. The selection of appropriate methods of data collection plays a vital role in ensuring accurate and meaningful results. Various factors need to be considered when determining the most suitable methods for collecting data. Here are some key considerations: Research Objectives: The first step in selecting appropriate data collection methods is to clearly define the research objectives. Different research questions may require different data collection techniques. For example, if the objective is to understand consumer preferences, surveys or interviews may be appropriate. If the objective is to analyze sales trends, collecting sales transaction data might be more suitable. Nature of the Data: Consider the type of data required for the analysis. Data can be quantitative (numeric) or qualitative (descriptive). Quantitative data often involves measurements, such as age, income, or product ratings. Qualitative data, on the other hand, focuses on subjective information, such as opinions, experiences, or narratives. Depending on the nature of the data, methods like surveys, experiments, observations, or interviews may be used. Population and Sample: Determine the population of interest and whether it is feasible to collect data from the entire population or a representative sample. If the population is large, a sample may be more practical. Sampling techniques like simple random sampling, stratified sampling, or cluster sampling can be employed. If the population is small, it may be possible to collect data from all individuals or units. Data Collection Instruments: Choose the appropriate tools or instruments for collecting data. Surveys, questionnaires, interviews, observation checklists, or measurement devices are common instruments. The selection depends on factors such as the type of data, the level of detail required, and the resources available. Ensure that the instruments are reliable, valid, and appropriate for the target population. Time and Resources: Consider the available time and resources for data collection. Some methods may be more time-consuming and costly than others. For instance, conducting face-toface interviews can be resource-intensive compared to online surveys. Evaluate the trade-offs between the desired level of accuracy and the practical constraints.

Ethical Considerations

Pay attention to ethical considerations when selecting data collection methods. Ensure that privacy and confidentiality are maintained, informed consent is obtained when necessary, and any potential risks to participants are minimized. Adhere to relevant ethical guidelines and regulations. Data Quality and Accuracy: Consider the potential for errors and biases associated with different data collection methods. Minimize sources of bias and ensure data quality through proper training of data collectors, clear instructions, and appropriate validation measures. Statistical techniques such as data cleaning and outlier detection can also be employed to enhance data accuracy. Feasibility and Practicality: Evaluate the feasibility and practicality of different data collection methods within the given research context. Consider factors such as the availability of participants, accessibility of the target population, logistical requirements, and budget constraints

Classification of data

In most of the researches huge amount of data are collected and is called raw data. By using the raw data we cannot predict or suggest anything about the research problem in focus. Hence, the raw data are grouped into different homogenous groups so that we can use it for the statistical analysis. This grouping of data is called as classification of data. The practice of grouping data based on shared features is known as classification. The data that have similar traits are grouped together into a class, which further divides the entire set of data into many groups or classes. For example human population can be classified into two broad groups on the basis of sex, males and females; or on the basis of literacy, literate and illiterate; or plants can be arranged into various classes according to their heights. The characteristics are of two types: descriptive and numerical. Descriptive: These characteristics include the individuals which cannot be numerically measured such as blindness, education, honesty, sex etc. Numerical: These characteristics include the individuals which can be numerically measured such as age, height, weight, yield etc.

1.4 Summary

Graphical representation refers to the use of visual elements such as charts, graphs, maps, and diagrams to represent data and information. These visual tools help to simplify complex data

sets, making them easier to understand and analyze. Here's a summary of the key types and purposes of graphical representation:

Types of Graphical Representation

1. Charts and Graphs:
    - Bar Chart: Used to compare quantities of different categories.
    - Line Graph: Displays data points over time, showing trends.
    - Pie Chart: Shows proportions and percentages among categories.
    - Histogram: Represents the distribution of numerical data.
    - Scatter Plot: Shows relationships between two variables.
2. Maps:
    - Geographical Maps: Visualize spatial distribution and patterns.
    - Heat Maps: Highlight areas of higher or lower intensity of data.
3. Diagrams:
    - Flowcharts: Depict processes or workflows.
    - Network Diagrams: Illustrate connections and relationships.
    - Tree Diagrams: Show hierarchical structures.
4. Tables: Organize data in rows and columns for easy comparison.

Purposes of Graphical Representation

1. Simplification: Condenses complex data into an understandable format.
2. Comparison: Makes it easy to compare different data sets or categories.
3. Trend Analysis: Helps identify trends and patterns over time.
4. Correlation: Illustrates relationships between different variables.
5. Distribution: Shows how data is spread across a range or area.
6. Decision Making: Assists in making informed decisions based on data insights.
7. Communication: Enhances presentations and reports by making data visually appealing and easier to grasp.

Benefits of Graphical Representation

1. Clarity: Improves clarity and comprehension of data.

2. Efficiency: Saves time by presenting data concisely.
3. Engagement: Engages the audience with visual appeal.
4. Insight: Provides deeper insights through visual analysis.
5. Overall, graphical representation is a powerful tool in data analysis and communication, making it easier to interpret and convey complex information effectively.

1.5 Keywords

- Data visualization
- Collection of secondary data
- Methods of data collection

1.6 Self-Assessment questions

1. What is the primary purpose of using graphical representations in biostatistics?
2. List three types of graphical representations commonly used in biostatistics.
3. How do histograms differ from bar charts in terms of their data representation?
4. When would you use a box plot in biostatistics? Provide an example.
5. Describe how scatter plots can be used to analyze relationships between two variables.
7. Blood pressure in a sample population. What variables would you include, and why?

1.7 Case Study

A public health researcher is conducting a study to understand the prevalence and distribution of diabetes in a small urban population. The researcher collects data on the age, gender, BMI, and diabetes status of 1,000 individuals. The goal is to present the data in a way that is easily understandable and reveals key insights.

Data Collected:

1. Age: Continuous variable
2. Gender: Categorical variable (Male, Female)
3. BMI: Continuous variable
4. Diabetes Status: Binary categorical variable (Diabetic, Non-diabetic)

1.8 References

1. Text Book of Biostatistics I. (2005). India: Discovery Publishing House Pvt. Limited.
2. Forthofer, R. N., Lee, E. S. (2014). Introduction to Biostatistics: A Guide to Design, Analysis, and Discovery. United States: Elsevier Science.

UNIT - 2

Univariate Graphical Representation

Learning Objectives

- Understand the Tabulation of data
- Understand the Graphical representation
- Understand the univariate presentation

Structure

2.1 Tabulation of data

After the classification, now the data are arranged in a concise and logical order, this process is known as tabulation of data. Tabulation is the technical term for the systematic grouping of data into rows and columns in a table. Tabulation of data is very significant because it helps in the comparison of data, detection of errors and omissions. Types of Tabulation: There are various kinds of tables are present but generally tabulation is of two types: Simple tabulation and complex tabulation. Simple Tabulation: This type of table provides information related to about one or more groups of independent questions. For instance, the population of plants of an area can be divided into many families. Here are only two columns, one for the different families and the other for the number of genera fall under that particular family. In this table we study only one point which is pertaining to the number of genera falling under various families and because of that reason it is also called as simple table or one-way table.

2.2 Graphical representation of the data

Data is visually expressed through charts and plots in a graphical representation. A graph can be used to visually represent the tabular data for more convenient analysis. Here, the data and statistical findings are shown visually. Comparing graphs to tables, it is much easier to interpret the findings. One extremely useful way to display the data is through graphics. The type of statistical findings and the characteristics of the data are taken into consideration when choosing from among the many different types of graphical representations, which include graphs, plots, charts, and diagrams. Listed below are a few frequently used graphical data representations:

Histogram

A vertical bar chart that displays the distribution of a collection of data is called a histogram. It is the most often used type of graphical data representation. It's employed to present and arrange the info in a format that's easier to navigate. It makes identifying the differences in the data quite simple. Assume we have twenty trees' worth of height data. How should we plot this data to create a histogram?

Graph paper should be used to draw a histogram. The values of the variable should be placed on the horizontal axis, or X-axis, and the frequencies should be placed on the vertical axis, or Y-

axis. A rectangle is produced for each class interval, with the height determined by the class interval's frequency and the base equal to the interval's length.

Bar diagram or bar graph or bar chart

A popular and understandable method of displaying categorical data or any ungrouped discrete frequency observations is through the use of bar charts.Let us suppose we have the data of mode of transport of a group of students (Table2.1) and we want to know which is the most and least used mode of transport by them to come to college.

Student	Mode	Student	Mode	Student	Mode
1	Car	11	Walk	21	Walk
2	Walk	12	Walk	22	Metro
3	Car	13	Metro	23	Car
4	Walk	14	Bus	24	Car
5	Bus	15	Train	25	Car
6	Metro	16	Bike	26	Bus
7	Car	17	Bus	27	Car
8	Bike	18	Bike	28	Walk
9	Walk	19	Bike	29	Car
10	Car	20	Metro	30	Car

Table 2.1 : Mode of transport of a group of students

So the first thing we will do here is to count the number of times a mode of transport comes (frequency) and make a list of it (Table 2.2), like the one mentioned below. After that, put these frequency values in a graph.

Mode	Frequency
Car	10
Walk	7
Bike	4
Bus	4
Metro	4
Train	1
Total	**30**

Table 2.2 : Frequency of mode of transport

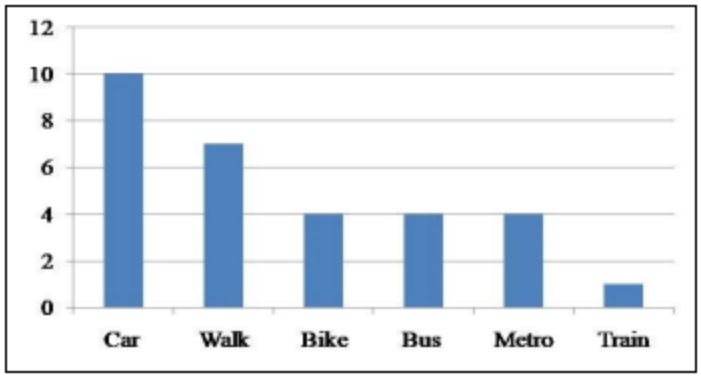

Figure:2.1. Bar graph showing various modes of transport

The vertical axis of the chart depicts a discrete number, while the horizontal axis displays the particular categories that are being compared. This bar graph makes it quite evident which vehicle is the most used form of transportation, with cars coming in first, followed by buses, bikes, and metro. This technique offers a quick and easy approach to identify basic popularity trends within a discrete data collection.

Frequency polygon

The frequency polygon resembles a histogram in many ways, but instead of using bars to represent each class, it uses single points that are connected by straight lines. Plotting data frequencies in various classes using frequency polygons is a good way to highlight patterns and trends in the data.

The polygon shown below (Figure 2) is based on the data used to draw the above mentioned histogram.

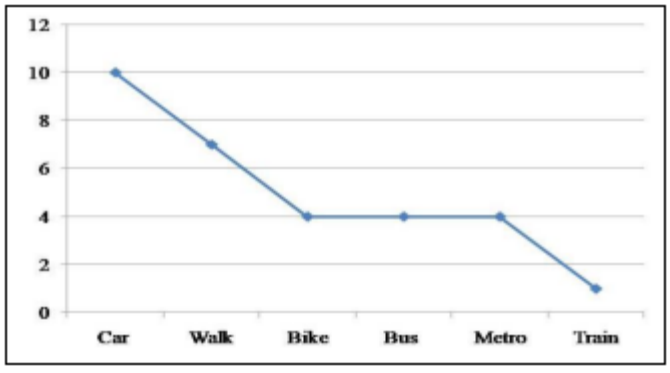

Figure: 2.2 : Showing the frequency polygon representation of dataCumulative frequency curve

The cumulative frequencies for the classes in a frequency distribution can be shown using the cumulative frequency curve. To put it another way, cumulative frequency is the total of all the frequencies that have been accumulated up to the distribution's upper class border. The number of data values that fall below a given boundary is shown. The class borders are labeled on the x (horizontal) and the cumulative frequencies are labeled on the y (vertical) axis before the curve is drawn. With the height representing the associated cumulative frequency, plot the cumulative frequency at each upper class border. Join the segments at the spots.

Pie chart

Pie chart is just a circle which is divided into segments. Each segment represents each value. It presents data, facts, and figures in an understandable "pie slice" style with different slice widths designating the quantity of each data item. It is very helpful to a layman to understand the statistics because anyone can measure the size of the slice of the cake or pie they are receiving whether it is small or large.

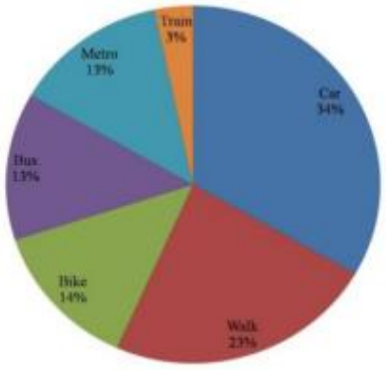

Figure: 2.3: Pie-chart showing the graphical representation of the data

Pictogram

In a pictogram instead of bars or charts symbols or pictures are used. It is used to show huge differences between categories.

2.3 Types of data for univariate presentation

Univariate data refers to data that consists of a single variable. When presenting univariate data, various types of data can be utilized. Here are some common types:

1. **Nominal Data**: This kind of data shows categories that aren't inherently ranked or ordered. Examples include marital status (single, married, divorced), gender (male, female), and automobile type (sedan, SUV, truck).
2. **Ordinal Data**: Ordinal data, unlike nominal data, have a natural order or ranking. However, the differences between the ranks are not quantifiable. Examples include ratings (1-star, 2-star, 3-star), educational levels (high school, bachelor's, master's), or satisfaction levels (low, medium, high).
3. **Interval Data**: Interval data lack a real zero point and have a discernible gap between values. One such example is temperature, expressed either in Celsius or Fahrenheit. Ratios have no interpretation in interval data. For instance, just because 20°C and 30°C differ by the same amount as 30°C and 40°C does not imply that 40°C is "twice as hot" as 20°C.
4. **Ratio Data**: All the features of interval data are included in ratio data, along with a real zero point that indicates the amount being measured's lack. Time, money, height, and weight are typical examples. Ratios have meaning in ratio statistics. Example, if one person's income is twofold another's, it means they have twofold as much.

When presenting univariate data, the choice of type depends on the nature of the variable being analyzed and the level of measurement it represents. Each type has its appropriate methods of presentation, such as frequency distributions, histograms, bar charts, pie charts, and summary statistics.

2.4 Summary

Tabulation and Univariate Presentation are fundamental techniques in biostatistics for organizing, summarizing, and interpreting data. Tabulation provides a structured format for data comparison, while univariate analysis helps in understanding the characteristics of a single variable through statistical measures and graphical tools. Together, these methods form the basis for more complex biostatistical analyses and help in making informed decisions based on the data.

2.5 Keywords

- Univariate Presentation

- Tabulation of Data
- Frequency polygon
- Pie Chart

2.6 Self-Assessment Questions

1. What is univariate presentation and how does it differ from bivariate and multivariate presentations?
2. List three common methods used in univariate presentation of data.
3. Explain what a frequency distribution is and provide an example using a dataset of your choice.
4. What is the difference between absolute frequency and relative frequency?
5. Define the mean, median, and mode. How are they useful in summarizing univariate data?
6. Given the following dataset: [3, 7, 7, 2, 5], calculate the mean, median, and mode.

2.7 Case Study

A biostatistician is analyzing the cholesterol levels (in mg/dL) of a group of 20 participants in a health study. The goal is to summarize the data using univariate analysis techniques and interpret the results to understand the distribution and central tendencies of cholesterol levels in the study group. Data Collected:

Cholesterol Levels (mg/dL): 190, 205, 189, 215, 220, 198, 210, 225, 230, 235, 245, 220, 210, 200, 195, 210, 205, 200, 215, 220

Also, Create a frequency distribution table for the cholesterol levels provided.

Cholesterol Level Range	Frequency
180-190	2
191-200	4
201-210	5
211-220	4
221-230	3
231-240	1
241-250	1

2.8 References

1. Text Book of Biostatistics I. (2005). India: Discovery Publishing House Pvt. Limited.
2. Forthofer, R. N., Lee, E. S. (2014). Introduction to Biostatistics: A Guide to Design, Analysis, and Discovery. United States: Elsevier Science.

UNIT - 3

Bivariate Graphical Representation

Learning Objectives

- Understand the bivariate presentation
- Understand the Scatter plots
- Understand the Line graphs and Bubble charts

Structure

3.1 Types of data for bivariate presentation

Bivariate data refers to data involving two variables. When presenting bivariate data, various types of data can be utilized depending on the nature of the relationship between the variables. Here are some common types:

1. **Nominal-Nominal**: In this type of bivariate data, both variables are nominal. Examples include analyzing the relationship between gender (male/female) and political affiliation (Democrat/Republican/Independent).
2. **Nominal-Ordinal**: One variable is nominal, and the other is ordinal. For instance, examining the relationship between job satisfaction levels (ordinal) and department names (nominal).
3. **Nominal-Interval/Ratio**: One variable is nominal, and the other is interval or ratio. An example could be studying the relationship between hair color (nominal) and IQ scores (interval/ratio).
4. **Ordinal-Ordinal**: Both variables are ordinal. This might involve exploring the relationship between educational attainment levels (ordinal) and income brackets (ordinal).
5. **Ordinal-Interval/Ratio**: One variable is ordinal, and the other is interval or ratio. For example, analyzing the relationship between performance ratings (ordinal) and sales figures (interval/ratio).
6. **Interval/Ratio-Interval/Ratio**: Both variables are interval or ratio. This includes scenarios like examining the relationship between temperature (interval/ratio) and ice cream sales (interval/ratio).

The choice of bivariate data type depends on the specific research question or hypothesis being investigated and the characteristics of the variables involved. Presentation methods for bivariate data include scatter plots, line graphs, contingency tables, correlation coefficients, and regression analysis, among others. These methods help to visually and analytically explore the relationship between the two variables.

3.2 Scatter plots, Line graphs and Bubble charts

Scatter diagram: The scatter diagram is used to the show the relationship between two variables and to prove or disprove cause-and-effect relationships. It is used to study the connections

between the two sets of data. For example the weight and height of a tree are related: the taller the tree the greater the weight. One variable is plotted on the horizontal axis and the other is plotted on the vertical axis. The graph below shows the positive correlation between the plant height and weight.

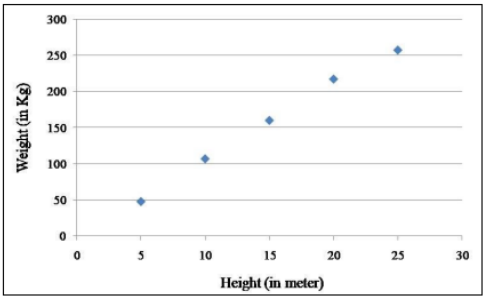

Fig 3.1 :Scatter diagram showing the representation of the data

Line graphs

The x and y axes of a line graph must match in scale for it to be helpful for comparing data sets. Since the values of the x-axis are independent of all other factors, it is sometimes referred to as the independent axis. Time, for instance, is always positioned on the x-axis since it is always changing, independent of other factors. Because the y-axis' values rely on factors on the x-axis—the company's revenues at this point—it is sometimes referred to as the dependent axis. As a result, there is a distinct value of y for every x value, and the straight line always advances horizontally.

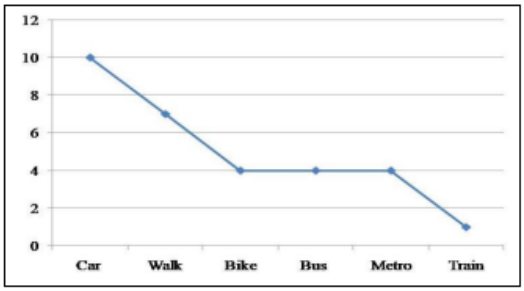

Fig 3.2 :Scatter diagram showing the representation of the data

Bubble charts

Creating a bubble chart involves representing data points with circles, where the size of each circle (bubble) corresponds to a third numerical variable

Example 1:

Let's consider a simple dataset of heights and weights of individuals:

Height (inches)	Weight (pounds)
65	150
70	160
68	155
72	180
66	145

Line Graph

Example 2:

For a line graph, we can use a dataset showing the monthly sales over a year:

Month	Sales (units)
January	50
February	45
March	60
April	70
May	80
June	90
July	100
August	110
September	95
October	85
November	75
December	65

Bubble Chart

Example 3:

For a bubble chart, let's consider a dataset with GDP, life expectancy, and population of different countries:

Country	GDP (Trillion $)	Life Expectancy (years)	Population (Million)
Country A	3.5	80	100
Country B	2.8	77	150
Country C	4.2	82	90
Country D	1.9	75	200
Country E	3.0	79	120

3.3 Summary

Bivariate presentation is crucial for understanding relationships between two variables, with various graphical methods available to visualize these relationships effectively:

- Scatter plots help identify correlations and patterns.
- Line graphs are excellent for showing trends over time.
- Bubble charts provide a way to visualize relationships between three variables simultaneously.

Each of these methods provides unique insights and helps in making data-driven decisions by clearly displaying the interactions between variables.

3.4 Keywords

- bivariate presentation
- Scatter plots
- Line graphs
- Bubble charts

3.5 Self-Assessment questions

1. What is bivariate presentation, and how does it differ from univariate presentation?
2. List three types of graphical representations commonly used in bivariate presentation.
3. Why is bivariate analysis important in biostatistics? Provide an example.
4. Describe the process of creating a bivariate table. What key elements should it include?
5. What trends can you identify from the line graph you created in question 13?
6. Create a bubble chart using the following data on GDP, life expectancy, and population size:

- GDP: [50, 55, 60, 70, 75, 80]
- Life Expectancy: [70, 72, 74, 76, 78, 80]
- Population Size: [10, 20, 30, 40, 50, 60]

3.6 Case Study

A researcher is conducting a study to examine the relationship between physical activity, diet, and health outcomes among adults. The dataset includes variables such as hours of physical activity per week, daily calorie intake, body mass index (BMI), and cholesterol levels. The researcher aims to use bivariate presentation techniques to explore the relationships between these variables.

3.7 References

1. Text Book of Biostatistics I. (2005). India: Discovery Publishing House Pvt. Limited.
2. Forthofer, R. N., Lee, E. S. (2014). Introduction to Biostatistics: A Guide to Design, Analysis, and Discovery. United States: Elsevier Science.

UNIT - 4

Measures of Central Tendency

Learning Objectives

- Understand the Central Tendency
- Understand the Mean, Median and Mode

Structure

4.1 Introduction

In this Unit we are going to study the concepts of biostatistics, first thing came in our mind is - what is biostatistics? Broadly speaking, biostatistics may be defined as the statistical techniques used to biological problem solving. The second question that arose was, "What are biological problems?" According to this definition, basic biological problems are those that arise in both basic biological sciences and applied fields like agriculture and health sciences. Before discussing more about biostatistics, we just recall what statistics is? The word Statistics was first time used by “Gattfried Ahenwall”. Generally speaking, the science of statistics is concerned with the gathering, categorization, analysis, and interpretation of numerical facts or data. According to Webster, “Statistics are the classified facts representing the conditions of the people in a state especially those facts which can be stated in number or in a table of number or in any tubular or classified arrangement” According to Bowley, “A numerical summary of the facts in any field of study arranged in relation to one another is called statistics. Furthermore, statistics is a science that deals with the gathering, presentation, analysis, and interpretation of numerical data, according to Croxton and Cowden.

This example makes it easy to understand how statistics are used. For instance, we keep hearing about the upcoming UP elections, and you occasionally see opinion polls. Based on these figures, what are the chances that Yogi Adityanath will defeat Akhilesh Yadav in the polls? Even with the facts that are regularly collected, you continue to find several polling businesses that produce disparate results. As a result, while one individual may predict that Yogi Adityanath will win by 50% of the vote, another may do so by 35%. Consequently, this suggests that this procedure is incredibly complex. While it is simple to calculate a number that indicates a difference, there is a highly active scientific process involved to ensure that the results are reliable. Now, let's define biostatistics. Thus, biostatistics is nothing more than the use of statistics to the study of biological processes, including those involving humans, animals, and living creatures. The use of statistics in Biology is known as biostatistics or biometery. Thus, we may draw examples from any discipline that deals with biology, medicine, bioengineering, or biology. For instance, if we want to talk about ecology, we can compare the structures of various sites to see how, over time, a mixed forest changed into a pine forest, and so on. To do these kinds of analyses, we must precisely analyze the structural elements of each individual forest component as well as any invasions that may have occurred over time, changes in topography,

and changes in climate. In microbiology lab, we want to predict the ability of synthesis drug to study the zone of inhibition to a different microbial strains, so on different bacterial we do the antibacterial assay and then in last came with the observation that how effective the synthesis drug is or how resistance the bacteria is for the drug and this open the aspect of market either to launch the particular drug in the market or not. For instance, in the last instance, everyone is aware of the harm the corona virus is now causing to public health. In light of this, anytime we travel to the airport or any other location, we will see that individuals arriving from various locations are required to have a necessary health examination in order to rule out the potential that they are corona virus carriers. These are instances of well-informed judgments made by the use of statistics: first, measures were taken, then, based on those measures, analyses were produced, and finally, based on those analyses, predictions were made on the appropriate course of action for correction. If in general we discuss the usefulness of biostatistics, we can say that it helps in presenting large quantity of data in a simple and more classified form which is easily understandable. It gave the methods of comparison of data and makes some predictions or judgments. Further, it finds out the possible relationship between the different variables.

4.2 Mean , Median and Mode

We usually encountered with large set of data's, because whatever our approach is for data collection, we have to collect data at least in replicates so that more reliable conclusion can be drawn. A single value is needed to describe the entire mass of unmanageable data. Therefore, biostatistics gives us the means to obtain a single value—a central value or an average—that more accurately characterizes the set of data. Or in other words we can say that, from the collected data the values of the variable tend to concentrate around some central value of observation, that value can be used as the representative value. So this tendency of distribution is known as central tendency. In terms of central tendency measurements most commonly studied averages are mean, median, and mode.

M**ean**

Arithmetic Mean Arithmetic Mean is the most popular and commonly used measure of central tendency. It is defined as the number obtained by dividing the total values of different items by

their number and it's denoted by . In general if arithmetic mean for ungrouped data or individual observation is, , , ,…… be 'n' observations for a variable x, the arithmetic mean is given by

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \cdots X_n}{n}$$

To further simplify the writing of a sum, the Greek letter∑ (sigma) is used. The sum x1 + x2 + x3 +⋯ xn is denoted,

$$\bar{X} = \sum_{i=1}^{n} Xi\,/n$$

Calculation of arithmetic mean

1. Series of individual observation
2. Discrete series
3. Continuous series

Series of individual observations: The calculation of arithmetic mean for such data which are present in individual series, the calculations is easy. Generally we have to get the total of values and divide this total by number of observation. Suppose we have five values for stomata's on leaves of a plant is 14 for first leaf, 17 for second leaf, 13, 15, and 19 for third, fourth and fifth leaf respectively. The arithmetic mean therefore is

$$\frac{14 + 17 + 13 + 15 + 19}{5} = 15.6$$

So the average number of stomata for the plant is 15.6

Symbolically

Leaf's	Stomata's
A	14
B	17
C	13
D	15
E	19
N=5	**ΣX= 15.6**

Further, the arithmetic mean for series of individual observation can be calculated by two methods:

A. Direct method

B. Assumed mean method or short cut method

A. Direct method

Example 1 : Calculate the arithmetic mean of the following marks in Hindi obtained by 10 students in a unit test.

Student	A	B	C	D	E	F	G	H	I	J
Marks	15	17	13	16	18	19	14	16	17	18

Wok Procedure

Obtained $\sum X$ by adding all the variables and divide the total by number of observation (N)

$\Sigma X=15+17+13+16+18+19+14+16+17+18$

$\Sigma X=163$

$N=10$

Thus, $\bar{X}= \Sigma X/N,$

$\bar{X}=163/10, 16.3$

The average marks in Hindi is 16.3

B. Assumed mean method or short cut method

Example 2 : Calculate the arithmetic mean of the following marks in Hindi obtained by 10 students in a unit test.

Student	Marks (X)	X-A (d)
A	15	-4
B	17	-2
C	13	-6
D	16	-3
E	18	-1
F	19	0
G	14	-5
H	16	-3
I	17	-2
J	18	-1
N=10		**-27**

Work procedure:

In this example, first we have to assume a mean, suppose assume mean=19. Calculate the deviation from assumed mean (X-A)=d

Get the total, of deviation from data using the following formula,

Mean $\bar{X}$= A+ Σd/N, where "A" is assumed mean and "d" is deviation

=19+(-27/10) = 16.3 marks

Thus the average marks in Hindi is 16.3

Discrete series

The value of each observation is multiplied by the frequency against each variable (observation) in a discrete series. Following this method, the values are totaled and divided by the entire number of frequencies. Symbolically,

$$\bar{X}=\frac{\Sigma fx}{\Sigma f}$$

Where, ΣfX = sum of the product of variables and frequencies.
Σf = sum of frequencies.

Example 3: The Number of seeds produced by 120 plants in garden given in table. Calculate the arithmetic mean.

Seed's (X)	Number of Plants (f)	fX
200	4	800
190	12	2280
180	15	2700
170	37	6290
160	22	3520
150	6	900
140	10	1400
	N=106	**Σf X= 17890**

Work Procedure

$$\bar{X}=\frac{\Sigma fx}{\Sigma f}$$

$$\bar{X}=\frac{17890}{106}$$

$$\bar{X}=168.77$$

Continuous series

Class intervals are provided when computing the arithmetic mean using this method. The process for computing the arithmetic mean in a continuous series is identical to that of a discrete

series. The sole distinction is the way the midpoints of different class intervals are acquired. Class intervals can be either inclusive or exclusive, or they can have different sizes, as we well know.

The following equation may be used to derived or get the mid-points.

$$\text{Mid-point (m)} = \frac{l1 + l2}{2}$$

here l_1 = lower limit and l_2 = upper limit

Example 4: Find out the mean of the following distribution in a class.

Marks	4-8	8-12	12-16	16-20
Student	4	8	6	3

Work Procedure

$$\text{Mid-point (m)} = \frac{l1 + l2}{2}$$

Marks (X)	Number of Student (f)	Mid- point	fm
4-8	4	6	24
8-12	8	10	80
12-16	6	14	84
16-20	3	18	54
	N = 21		**Σfm** = 242

$$\bar{X} = \frac{242}{21}$$

$$\bar{X} = 11.52$$

Some other types of mean's

1. **Geometric Mean (GM)**

GM of N variate value is the Nth root of their multiply. In algebra, geometric mean is used or calculated in case of geometric progression. Further, like arithmetic mean it also depends on all observations. Symbolically G.M = $(x_1.x_2\ldots\ldots\ldots\ldots.x_n)/N$ or $(x_1f_1.\ x_2f_2\ldots\ldots\ldots x_nf_n)/N$

2. **Harmonic Mean (HM)**

HM is described as reciprocal of AM the reciprocal of individual observations. In algebra, HM is found out in the case of harmonic progression only, but in statistics this mean is

suitable measure of central tendency (data pertains to time, speed and rates). This mean is rigidly defined and calculation is based on all the observations. Symbolically harmonic Mean is

$$H = \frac{N}{f1X1+F2X2+\cdots\ldots\ldots\ldots FkXk}$$

When calculating the arithmetic mean of this type of data, weights have been applied to different components based on their relative relevance. As an illustration, we have two commodities: potatoes and apples. Finding the average cost of these goods is of importance to us. Thus for the calculation of arithmetic mean we consider the condition in this way i.e., we consider P1 and P2 to our commodities and use the formula $P_1+P_2/2$. To put it another way, the basic arithmetic mean assigns each item a weight of equal significance, whereas in most cases, the elements in a series have varying degrees of value. However, in order to assign significance to the increase in potato prices (P2), we must first use the shares of apples and potatoes in the consumer's budget (W1) and W2, respectively, as "weights." Now the AM weighted by the shares in the budget known by next formula,

$$\frac{W1P1+W2P2}{W1+W2}$$

In general the weighted arithmetic mean is given by the following formula

$$W_1P_1\frac{W1P1+W2P2+W3p3+\cdots\ldots\ldots\ldots WnPn}{W1+W2+W3+\cdots\ldots\ldots\ldots Wn} = \frac{\Sigma WX}{\Sigma W}$$

Study of Median

Median is the positional value that divide the given observations or data into equal half, one part of the observation includes all value that are greater than the median and other part includes the values that are lesser than the observations. Furthermore, the term "median" refers to the central or middle value of a variable in a series of data, indicated by a capital "M," when the observations are ordered in either ascending or descending order of magnitude. In comparison to the arithmetic mean median is position average while the arithmetic mean is mathematical average because suppose we have a series of 7 observations i.e., 20,30,40,50,60,70 and 80, so from the observations 50 will be the median and in second condition if we change the observation of the series and we have new observation such as 10,30,40,50,60,70 and 80, but for

this series the median again comes to be 50. However, in the case of the arithmetic mean, a change in a single item affects the average value; in contrast, the median remains constant if items other than the center value change. Thus, this further confirms that the median is positional value.

Calculation of Median: As we encounter different observation or data's we have different working formulas for calculation of median

A. When the data is ungrouped data or for simple series:

Procedure

1) Arrange the "n" (number of values) values of the variable in either ascending or descending order of magnitudes.

2) When "n" is odd or the number of observations is odd, the $\frac{n+1}{2}\ th$ value is the median

Thus, M = $\frac{n+1}{2}\ th$ term

3) When "n" is even, then there are two value in the middle, so the median can be estimated by finding arithmetic mean of middle two values i.e. adding two values in the middle and dividing by two.

Thus, 'n' is even. In this case there are two middle terms $\frac{n}{2}\ th$ and $\left(\frac{n}{2}+1\right) th$.

So median will be calculate by using this formula

$$M = \frac{\frac{n}{2}+\left(\frac{n}{2}+1\right)}{2}$$

Example 5: The average weight of 9 students is:45, 48, 53, 46, 54, 59, 42, 48, 41 find the median.

Work Procedure

Arrange in ascending order

41, 42, 45, 46, 48, 53, 54, 58, 59

$$\text{Median} = M = \left(\frac{n}{2}+1\right) th$$

M = (9+1)/2 = 5^{th} value

Therefore the median is 48

B. When the data is grouped

1 **When the data is in discrete series:** In case of discrete series the position of median i.e. $(N+1)/2$ th item can be located through cumulative frequency.

Example 6: Find the median for the following data obtained for monthly salaries of peoples

No. of Persons	3	6	7	15	11	9	5	4	11	2
Income (in Rs.)	1300	700	1900	2100	2300	2500	6000	4800	3500	4000

Work Procedure

$$M = \left(\frac{n+1}{2}\right) th$$

Income (in Rs.)	No. of Persons	cumulative frequency
1300	3	3
700	6	9
1900	7	16
2100	15	31
2300	11	42
2500	9	51
6000	5	56
4800	4	60
3500	11	71
4000	2	73
	N = Σf = 73	

Here n= 73

M = (73+1)/2 th term

M = 37th value

Therefore the median is 2300

2 **When the series is continuous:** In case of continuous series you have to locate the median class where $N/2$ th item [not $(N+1)/2$ th item] lies. The median can then be obtained by using the following formula.

$$M = L + \frac{\frac{N}{2} - cf}{fm} \times i$$

Where, L = lower limit of class in which the median lies.

N = Total no of frequencies, n = Σf.

fm = frequency of the class in which the median lies.

C = cumulative frequency of the class proceeding to the median class

i = width of the class interval of the class in which the median lies.

Example 7: Find the median of the following data, the table is given the marks obtained by students in a Gymnosperm class.

Marks	20-25	25-30	30-35	35-40	40-45	45-50	50-55	55-60
Number of students	14	26	23	33	35	13	8	15

Work procedure:

1. A frequency table with a class interval is used to present the data in this instance.
2. In this case, the data is presented as a frequency table with a class interval.We determine cumulative frequencies for every value.
3. For every value, cumulative frequencies are discovered.Find the class where the median is located. N^{th} item is located in the class known as the median.
4. The class where the N^{th} item is located is known as the median class
5. The following formula can be used to get the precise value of the median after the Median Class has been established.

$$M = L + \frac{\frac{N}{2} - cf}{fm} \times i$$

Marks	Number of students	Cumulative frequency
20-25	14	14
25-30	26	40
30-35	23	63
35-40	33	96
40-45	35	131
45-50	13	144
50-55	8	152
55-60	16	168

Here, n=168, so n/2= 84

Thus, median class is 35- 40

Lower limit of the median class is 35.

Cumulative frequency of the class preceding the median class = 63.

Fm = median class = 33.

I class interval of the median class= 5.

Thus, putting all these values in given formula

$$M = L + \frac{\frac{N}{2} - cf}{fm} \times i$$

M = 35+5 = 40.

So the calculated Median for the given data is M= 40.

Thus, this mean half of the class have score equal or less than 40 marks in subject and half of the class have score higher than 40 marks in the subject.

Advantages (merits) of median

a. It is easily understood, easy to locate without any difficulty although it is not as popular as arithmetic mean.

b. The value is not affected by magnitude of extreme deviations.

c. Median can also be determined graphically to ogives.

d. It is very useful in open ended classes or where the extreme classes are ill defined, like "less than 20" or "more than 20".

e. Median is very good in case of qualitative data's or when the items are not susceptible to measurements in definite units.

f. Median is unaffected by abnormal value.

g. Median always remain the same whatsoever methods of computation be applied.

Study of Mode

Mode is generally defined as the most common or value in a series which appears most frequently. Because it appears the most frequently in the series, the French phrase "la Mode," which denotes the most stylish values of a distribution, is where the word "mode" originates. For example in a series 9, 8, 6, 9, 5, 6, 3, 9, 2, 1, 7, 9, 9, 4, 1, 9, 9, 8 we noticed that 9 comes seven time so the mode for the series is 9 or in other word, mode represent the maximum demanding figure. In other words, mode represents that value which is most frequent or typical or predominant. According to Croxton and Cowden, "The mode is that value or point around which the items to be most heavily concentrated." Further, according to Kenny and Reepura “the value of the variable which occurs most frequently in a distribution is called mode”. Mode is some time also known as Norm and denoted by Mo. Calculation of mode A. When data is simple or mode in simple series. In case of simple series, the value which appears in maximum number of time is mode of that series. By counting the number of times that the different values repeat themselves, the inspection technique may be used to identify it. The value that occurs the most number of times is the modal value.

Example 8: Mark of the 10 students are recorded as 26, 22, 37, 30, 45, 40, 37, 30, 37, 26, 30, 37, 45, 37, 22. Calculate the mode of the series.

Work procedure

1. Arrange the data in series and locate the value which occurs maximum number of times.
2. Write the number of time the value located in the data against each value.
3. The value comes maximum will be the mode of the series.

Marks Obtained	Number of time repeat
22	2
26	2
30	3
37	5
40	1
45	2
	ΣN = 15

The number 37 occurs for the largest number of times. So 37 is the mode of the above series.

Inspection Method

Example 9: Protein content of 15 milk samples was recorded as 18, 15, 8, 12, 6, 20, 9, 14, 12, 7, 12, 20, 19, 18, 9. Work procedure: Arrange the data in ascending order and then convert the observation into frequency distribution table.

Protein %	6	7	8	9	12	14	15	18	19	20
Frequency	1	1	1	2	3	1	1	2	1	2

During the inspection of the above table we find that 12 repeated three times, which is maximum from the given data. Therefore, the mode of the given observation is 12.

Grouping Method: When the discrete series is bimodal or multimodal then grouping method is used to obtained mode for the series.

Example 10: In an ecology survey data regarding girth of trees has been collected, calculate the mode from the following frequency distribution

Girth	110	115	120	125	130	135	140	145	150	155	160	165	170	175
frequecny	3	6	9	10	13	15	16	16	12	14	10	8	5	4

Work procedure

1. Arrange the value in ascending or descending orders.
2. Draw a table and arranged data in tabular form.
3. In column I put the values of variables
4. In column II frequency to be write against their values.
5. Sum of 2-2 frequency to be written in IIIrd column.
6. Again write sum of 2-2 frequencies in column IVth, ignoring the first frequency.
7. Further, repeat the process, but this time sum of 3-3 frequencies have to take and write in the column Vth
8. Again write sum of 3-3 frequencies in column VIth, ignoring the first and second frequency or in other words first two frequencies.
9. Repeat according to the need of the data for more grouping i.e., 4-4, 5-5- and 6-6
10. Detection of maximum frequency done in each column of group and variable noted in analysis table and variable with maximum repetition consider as mode.

Girth	Frequencies					
	Individual	Grouping by twos		Grouping by threes		
	I	II	III	IV	V	VI
110	3					
115	6	} 9				
120	9		} 15	} 18		
125	10	} 19			} 25	
130	13		} 23			} 32
135	15	} 28		} 38		
140	16		} 31		} 44	
145	16	} 32				} 47
150	12		} 28	} 44		
155	14	} 26			} 42	
160	10		} 24			} 36
165	8	} 18		} 32		
170	5		} 13		} 23	
175	4	} 9				} 17

Column	Girth of tree with maximum frequency
I	140, 145
II	140, 145
III	135, 140
IV	140, 145, 150
V	130, 135, 140
VI	135, 140, 145

So from the above results, we have reported that 140 occurs maximum number of time i.e., 6 times. Therefore mode is 140. 2.

Mode of a Continues Series In case, where we have data in continuous series we have to find out the class in which the mode is situated and the class in which mode is situated in known as modal class. After determining the modal class we calculate the mode by using the following formula.

$$\text{Mode} = L_1 + \left(\frac{D1}{D1+D2}\right) \times i$$

Where L_1 = Lower limit of modal class.

D1= difference between the frequency of the modal class and the frequency of the class preceding the modal class.

D2 = difference between the frequency of the modal class and the frequency of the class succeeding the modal class.

i= Class interval or width of the class

Or some time, another expression also used for computing the mode is

$$\text{Mode } (M_o) = L_1 + \left(\frac{f_m - f_1}{2f_m - (f_1 - f_2)}\right) \times i$$

Where L_1 = Lower limit of modal class

f_m = Frequency of modal class or the maximum frequency.

f_i= Frequency of pre-modal class.

f_2= Frequency of post modal class.

i = Class interval or class width of the class.

Example 11: Find the mode for the following distribution

Marks	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Number of students	6	9	8	12	17	7	1

Work Procedure

$$\text{Mode} = L_1 + \left(\frac{D1}{D1+D2}\right) \times i$$

where

D1 = 17-12 = 5

D2 = 17-10 = 7

$L_1 = 70$

i = 10

So, Mode = 70+3.33 = 73.33

Marks	No. of students
30—40	6
40—50	9
50—60	8
60—70	12
70—80	17
80—90	7
90—100	1

Formula 2, $M_o = L_1 + \left(\frac{f_m - f_1}{2fm-(f_1-f_2)}\right) \times i$

Here, L_1 is 70.

$F_m = 17$

$F_1 = 12$

$F_2 = 7$

$= 70 + \left(\frac{17-12}{2\times17-(12-7)}\right) \times 10$

$= 70 + \left(\frac{5}{34-19}\right) \times 10$

$= 70 + \frac{5}{15} \times 10$

$= 70 + 3.33$

$= 73.33$

So, the calculated mode for the given data is $M_o = 73.33$

Thus, results for both formula's remains same.

4.3 Summary

Comprehending the mean, mode, and median facilitates an exhaustive examination of the data, with each offering distinct perspectives. The type of data and the precise objectives of the research will determine which central tendency measure is best.

4.4 Keywords

- Central Tendency
- Mean
- Mode
- Median

4.5 Self-Assessment Questions

1. Define central tendency and explain its significance in statistical analysis.
2. Compare and contrast the mean, mode, and median. In what situations would you prefer one measure over the others?
3. Calculate the mean of the following dataset: [10, 15, 20, 25, 30].
4. Explain how the mean can be affected by skewed distributions.
5. Identify the mode(s) in the following dataset: [5, 5, 10, 15, 15, 20, 20].

6. Find the median of the following dataset: [12, 15, 18, 20, 25, 30, 35].
7. Explain how the median is less influenced by outliers compared to the mean.

4.6 Case Study

A human resources department is analyzing the salaries of employees in a company. The dataset contains salary information for 50 employees. The HR team wants to use central tendency measures to better understand the salary distribution and make informed decisions regarding compensation policies.

Data Collected:

Dataset: [30000, 35000, 40000, 45000, 50000, 55000, 60000, 65000, 70000, 75000, 80000, 85000, 90000, 95000, 100000, 105000, 110000, 115000, 120000, 125000, 130000, 135000, 140000, 145000, 150000, 155000, 160000, 165000, 170000, 175000, 180000, 185000, 190000, 195000, 200000, 205000, 210000, 215000, 220000, 225000, 230000, 235000, 240000, 245000, 250000, 255000, 260000, 265000, 270000, 275000]

Questions for Analysis:

1. Calculate the mean, mode, and median of the salary dataset.
2. Discuss any differences observed between the measures of central tendency.
3. How would you interpret the mode in this salary dataset? What does it indicate about the salary distribution?
4. Explain how you would use the median to determine a fair salary range for new hires in the company.
5. Discuss any potential limitations or biases in using the mean salary as a benchmark for compensation decisions.

4.7 References

1. Text Book of Biostatistics I. (2005). India: Discovery Publishing House Pvt. Limited.
2. Forthofer, R. N., Lee, E. S. (2014). Introduction to Biostatistics: A Guide to Design, Analysis, and Discovery. United States: Elsevier Science.

UNIT - 5

Measures of Location and Measures of Dispersion

Learning Objectives

- Understand the Measures of Location
- Understand the Measures of Dispersion

Structure

5.1 Quartiles, Quintiles, Deciles and Percentiles

Quartiles

Quartiles are a type of quantile and are values that divide your data into quarters, each containing 25% of the data. There are three quartiles: the first quartile (Q1), the second quartile (Q2 or the median), and the third quartile (Q3).

Here's a detailed breakdown of the quartiles:

First Quartile (Q1): Also known as the lower quartile, it is the median of the lower half of the dataset (25th percentile). It separates the lowest 25% of the data from the rest.

Second Quartile (Q2): This is the median of the dataset (50th percentile), separating the dataset into two equal halves.

Third Quartile (Q3): Also known as the upper quartile, it is the median of the upper half of the dataset (75th percentile). It separates the highest 25% of the data from the rest.

Interquartile Range (IQR): This is the range between the first and third quartiles (Q3 - Q1). It measures the spread of the middle 50% of the data.

Quintiles

- Quintiles divide a dataset into five equal parts.
- For example, the second quintile (Q2) represents the median, while the first quintile (Q1) and the third quintile (Q3) divide the dataset into two equal parts each.

Deciles

Deciles are another form of quantile, dividing a dataset into ten equal parts. Each decile represents 10% of the sorted dataset, and there are nine deciles in total. Here’s a breakdown of how they are defined:

D1 (First Decile): 10th percentile of the data.

D2 (Second Decile): 20th percentile of the data.

D3 (Third Decile): 30th percentile of the data.

D4 (Fourth Decile): 40th percentile of the data.

D5 (Fifth Decile): 50th percentile (also the median).

D6 (Sixth Decile): 60th percentile of the data.

D7 (Seventh Decile): 70th percentile of the data.

D8 (Eighth Decile): 80th percentile of the data.

D9 (Ninth Decile): 90th percentile of the data.

To calculate the deciles for a given dataset, you first need to sort the data and then find the values that correspond to the 10th, 20th, ..., 90th percentiles.

Percentiles

- Percentiles divide a dataset into 100 equal parts.
- For example, the 25^{th} percentile signifies the Q1, the 50^{th} percentile signifies the median, and the 75^{th} percentile represents the Q3.

5.2 Mean deviation and Quartile deviation

Study of Mean Deviation

The arithmetic mean of the deviations of different items from a central tendency measure is called the mean deviation of a series. Although, mathematically, mean deviation is not a logical measure of dispersion more this method in not valid for algebraic expressions. Calculation of mean deviation: Like other measures, mean deviation is also calculated for all three types of datas.

- Series of individual observation.
- Discrete series.
- Continuous series.

A. Series of individual observation:

Example 1: In a gymnosperms exam the marks obtained by the ten students is given, calculate mean deviation and its coefficient from the data.

Marks	62	68	66	79	87	75	60	89	77	83

Work procedure

Calculate the mean deviation using formula, M.D. $= \frac{\Sigma |D|}{N}$

Then calculate the coefficient of mean deviation using the following formula,

Coefficient of M.D $= \frac{M.D}{Mean}$

Marks	Deviation from Mean (ignoring signs) $\|D\|$
62	12.6
68	6.6
66	8.6
79	4.4
87	12.4
75	0.4
60	14.6
89	14.4
77	2.4
83	8.4
$\Sigma X = 746$	$\Sigma \|D\| = 84.8$

Step I: Calculate the arithmetic mean Here we have,

$\Sigma X = 746$ and $N = 10$

Arithmetic mean $= \bar{X} = \Sigma X/N$

$\bar{X} = 746/10$

$= 74.6.$

Step II: Calculate the deviation, ignoring signs

Calculate deviation is $\Sigma |D| = 84.8$

Step III: calculate mean deviation

M.D. $= \frac{\Sigma |D|}{N}$

M.D $= 84.8/10, = 8.48$

Thus, M.D is 8.48.

Step IV: Calculate coefficient of Mean deviation

Coefficient of M.D $= \frac{M.D}{Mean}$

$= 8.48/74.6$

$= 0.113$

In conclusion we have, Arithmetic mean = 74.6; Mean deviation = 8.48 and Coefficient of mean deviation from the data = 0.113.

Example 2: Calculate mean deviation and its coefficient from the data.

No of accidents	0	1	2	3	4	5	6	7	8	9	10	11
Person involved	12	8	15	21	6	15	14	11	2	1	0	2

Work procedure

Arrange the data accordingly into the table and then calculate its cumulative frequency.

Calculate the median using $\frac{N+1}{2}$

Calculate the deviation from median ignoring signs.

Calculate the mean deviation using formula, M.D. $=\frac{\Sigma f \mid D \mid}{N}$

Then calculate the coefficient of mean deviation using the following formula,

Coefficient of M.D $= \frac{M.D}{Mean}$

Number of accidents	Person involved	Cumulative frequency c.f.	|D|	f |D|
0	12	12	3	36
1	8	20	2	16
2	15	35	1	15
3	21	56	0	0
4	6	62	1	6
5	15	77	2	30
6	14	91	3	42
7	11	102	4	44
8	2	104	5	10
9	1	105	6	6
10	0	105	7	0
11	2	107	8	16
	N= 107			Σ f |D| =221

Step I: Calculate cumulative frequency

Step II: Calculate median using $\frac{N+1}{2}$

= 107+1/2

= 54th term or item, since for the 54th item or in other words value located at the size of the item in which cumulative frequency falls is 3.

Therefore, Median is 3

Step III: Calculate the deviation using median.

Step IV: Calculate mean deviation

M.D. $=\frac{\Sigma f \mid D \mid}{N}$

M.D = 221/107, = 2.06

Thus, M.D is 2.06

Step IV = Calculate coefficient of Mean deviation

Coefficient of M.D $= \frac{M.D}{Median}$

= 2.06/3

= 0.68

In conclusion we have, Median = 3; Mean deviation = 2.06 and Coefficient of mean deviation from the data = 0.68.

Quartile deviation

One way to quantify statistical dispersion and show the spread or variability within a dataset is to use the quartile deviation. It is based on quartiles, which split a dataset into four equal sections, each holding twenty-five percent of the total data.
Finding the first quartile (Q1) and the third quartile (Q3) is the first step towards calculating the quartile deviation. Once you have these numbers, you may use the formula to get the quartile deviation.

$$QD=Q3-Q1$$

1. **Find Q1 and Q3**: To catchQ1 andQ3, you necessity to assemble the data in arising order and then trace the values that mark off the first 25% and the third 25% of the data, respectively.
2. **Calculate the quartile deviation (QD)**: Subtract Q1 from Q3, and then divide the result by 2. This gives you the quartile deviation, which represents the spread of the middle 50% of the data.

The quartile deviation is often used as a measure of variability in skewed distributions or those with outliers because it is less affected by extreme values compared to the range or standard deviation. However, it provides less information about the distribution shape compared to other measures like the standard deviation.

5.3 Variance

A statistical metric called variance is used to quantify how far apart or dispersed a group of data points are. It is a measurement of the degree to which the mean (average) value of a dataset deviates from the individual values within the dataset.

To compute the variance, you track these steps:

1. **Compute the Mean**: Add up all the data points and divide by the total number of data points. This gives you the mean (average) value.
2. **Compute the Squared Differences**: For each data point, subtract the mean from that data point and square the result. This ensures that negative differences don't cancel out positive differences.
3. **Calculate the Variance**: Find the average of all the squared differences calculated in step 2.

The formula for variance σ^2 of a population is:

$\sigma^2=\sum(xi-\mu)^2$/n-1

Where:

- xi represents all individual data point.
- μrepresents the mean of the dataset.
- N is the total sum of data points.

If you're dealing with a sample rather than an entire population, you use a slightly different formula for the sample variance s^2:

$$s^2=\sum(xi-x^-)2/n-1$$

Where:

- xirepresents all individual data point.
- x^- represents the sample mean.
- $n$is the total sum of data points in the sample.

Variance is useful because it provides a measure of the variability or dispersion of data points around the mean. However, because variance is in squared units (e.g., squared dollars, squared meters), it's often

5.4 Summary

Descriptive statistics measures such as quartiles, quintiles, deciles, percentiles, quartile deviation, mean deviation, and variance provide valuable insights into the distribution and variability of data. By understanding these measures and their calculations, analysts can better interpret and summarize datasets, leading to informed decision-making processes.

5.5 Keywords

- Quartiles and Quintiles
- Deciles and Percentiles
- Quartile deviation
- Mean deviation
- Variance

5.6 Self-Assessment questions

1. Define quartiles and explain their significance in data analysis.

2. Calculate the quartiles for the following dataset: [10, 15, 20, 25, 30, 35, 40, 45, 50].
3. Calculate the 20th percentile for the following dataset: [5, 10, 15, 20, 25, 30, 35, 40, 45, 50].
4. Calculate the quartile deviation for the following dataset: [10, 15, 20, 25, 30, 35, 40, 45, 50].
5. Calculate the mean deviation for the following dataset: [10, 15, 20, 25, 30, 35, 40, 45, 50].
6. Calculate the variance for the following dataset: [10, 15, 20, 25, 30, 35, 40, 45, 50].

5.7 Case Study

A teacher wants to analyze the test scores of her students to understand the distribution and variability of their performance. The dataset contains the test scores of 50 students.

Data Collected:

Test Scores: [65, 70, 75, 80, 85, 90, 95, 100, 105, 110, 115, 120, 125, 130, 135, 140, 145, 150, 155, 160, 165, 170, 175, 180, 185, 190, 195, 200, 205, 210, 215, 220, 225, 230, 235, 240, 245, 250, 255, 260, 265, 270, 275, 280, 285, 290, 295, 300]

Questions :

1. Calculate the quartiles and quintiles for the test scores dataset.
2. Find the 30th and 70th percentiles for the test scores dataset.
3. Determine the quartile deviation, mean deviation, and variance for the test scores dataset.
4. Discuss the insights gained from analyzing the quartiles, quintiles, and dispersion measures.
5. Based on the analysis, what recommendations would you make to the teacher regarding student performance?

5.8 References

1. Text Book of Biostatistics I. (2005). India: Discovery Publishing House Pvt. Limited.
2. Forthofer, R. N., Lee, E. S. (2014). Introduction to Biostatistics: A Guide to Design, Analysis, and Discovery. United States: Elsevier Science.

UNIT - 6

Introduction to Standard Deviation

Learning Objectives

- Understand the Standard Deviation
- Understand the Actual Mean Method
- Understand the Assume Mean Method

Structure

6.1 Definition and importance

Karl Person applied the idea of standard deviation for the first time in 1823 and today its most unusually used amount of dispersion in statistics work and frequently used in biological samples and satisfies most of the characteristics of good measure. Standard deviation is denoted by Greek letter σ (sigma) and abbreviated variously as S.D. or SD. By definition standard of deviation is defined as "Square root of the arithmetic average of the squares of the deviations measured from the mean". Moreover, the majority of research reports the SD, which is an indication of the original data points' variability.

In general it is figured by six general steps

1. Compute the mean.
2. Determine how each observation differs from the mean.
3. Square any deviations between the data and the mean.
4. To obtain the sum of squares for the deviation, add the squared values.
5. This total may be divided by the number of observations minus one to obtain variance (σ), also known as mean-squared deviation.
6. To obtain the root-mean squared deviation, often known as the standard deviation, find the square root of this variance. After the original has been squared, take the square root in the opposite direction.

Calculation of standard deviation: For calculation we may counter all three types of data i.e.,

A. Series of individual observation.

B. Discrete series.

C. Continuous series.

6.2 Calculation methods

A. Series of individual observation.

When data is in series of individual observations: in case when data is in series of individual method we may use both (i) Actual mean method and (ii) assumed mean method

Example 1: In a class of 10 students, compute the standard deviation of their marks in Botany paper.

45, 70, 65, 84, 72, 68, 91, 59, 77, 89.

Work procedure:

Calculate the actual mean of the observations.

Calculate the deviation (x) of the value (marks) from the mean (X-$\bar{X}$).

Square the deviation and calculate the Σx^2.

Calculate the Standard deviation using the following formula

$$SD\ (\sigma) = \sqrt{\frac{\Sigma x^2}{N}}$$

1. By Actual mean Method:

Step 1: calculate the mean from the observations

Here, x = (X-$\bar{X}$)

For the calculation of $\bar{X}$ we know that $\bar{X} = \Sigma X/N$

i.e., $\bar{X} = 720/10$

So the actual mean of the observation is $\bar{X} = 72$

Step II: calculate the standard deviation by using the formula

$$SD\ (\sigma) = \sqrt{\frac{\Sigma x^2}{N}}$$

$$= \sqrt{\frac{1786}{10}}$$

$$= \sqrt{178.6}$$

So the deviation of marks among the ten student is, $\sigma = 13.36$.

2. By Assumed mean Method

Marks (X)	x = (X-68)	x^2
45	-23	529
70	2	4
65	-3	9
84	16	256
72	4	16
68	0	0
91	23	529
59	-9	81
77	9	81
89	21	441
	$\Sigma x = 40$	**$\Sigma x^2 = 1946$**

Here we have assumed 68 as mean from the observation.

Formula for standard deviation

$$\sigma = \sqrt{\frac{\sum x^2}{N} - \frac{(x)^2}{N}}$$

here, $\Sigma x^2 = 1946$; $\Sigma x = 40$ and $N = 10$

so put all these value in the formula and we get

$$\sigma = \sqrt{\frac{\sum 1946^2}{10} - \frac{(40)^2}{10}}$$

$$\sigma = \sqrt{194.6 - 16}$$

So the deviation of marks among the ten student is same by both methods i.e., $\sigma = 13.36$.

Example 2: In a survey of ten villages for graduate students data was recorded (X= 36, 49,79, 51, 37, 87, 63, 74, 31, 43), calculate the Standard deviation from the observation.

1. By actual mean method.

Graduate people (X)	$x = (X-\bar{X})$	x^2
36	-19	361
49	-6	36
79	24	576
51	-4	16
37	-18	324
87	32	1024
63	8	64
74	19	361
31	-24	576
43	-12	144
$\Sigma X = 550$		$\Sigma x^2 = 3482$

Work Procedure

Step 1: calculate the mean from the observations Here, x = = $(X-\bar{X})$

For the calculation of $\bar{X}$ we know that $\bar{X} = \Sigma X/N$ i.e., X = 550/10

So the actual mean of the observation is $\bar{X} = 55$

Step 2: calculate the standard deviation by using the formula

$$SD\ (\sigma) = \sqrt{\frac{\sum x^2}{N}}$$

$$= \sqrt{\frac{3482}{10}}$$

$$= \sqrt{348.2}$$

So the deviation of marks among the ten student is, $\sigma = 18.66$.

2. By assumed mean method.

Graduate people (X)	x = (X-51)	x^2
36	-15	225
49	-2	4
79	28	784
51	0	0
37	-14	196
87	36	1296
63	12	144
74	23	529
31	-20	400
43	-8	64
	$\Sigma x = 40$	$\Sigma x^2 = 3642$

Here we have assumed 68 as mean from the observation.
Formula for standard deviation

$$\sigma = \sqrt{\frac{\Sigma x^2}{N} - \frac{(x)^2}{N}}$$

here, $\Sigma x^2 = 3642$; $\Sigma x = 40$ and $N = 10$
so put all these value in the formula and we get

$$\sigma = \sqrt{\frac{\Sigma 3642^2}{10} - \frac{(40)^2}{10}}$$

$$\sigma = \sqrt{364.2 - 16} \qquad = \sqrt{348.2}$$

So the deviation of marks among the ten student is same by both methods i.e., $\sigma = 13.36$.

B. For Discrete Series:

Example 3: In a sample from pond we got fishes of different weight, calculate the standard deviation within the observation.

Weight of fishes (Kg)	1	2	3	4	5
Frequency	13	12	10	6	4

In case of discrete method standard deviation can be calculated by both methods
Actual mean method and assumed mean method

1. Actual mean method

Work procedure:

- Calculate the actual mean of the observations.
- Calculate the deviation (x) of the value from the mean $(X-\bar{X})$.
- Square the deviation and calculate the Σx^2

- Multiple them by the respective frequencies and make the total i.e., Σfx^2
- Calculate the Standard deviation using the following formula.

$$SD\ (\sigma) = \sqrt{\frac{\Sigma fx^2}{N}}$$

Weight (X)	Frequency (f)	fX	$x = (X-\bar{X})$	X^2	fx^2
1	13	13	-1.46	2.16	28.08
2	12	24	-0.46	0.22	2.64
3	10	30	0.53	0.28	2.8
4	6	24	1.53	2.34	20.34
5	4	20	2.53	6.4	25.6
	N = 45	Σfx = 111			$\Sigma fx^2 = 79.46$

Here, Here, $x = (X-\bar{X})$
For the calculation of $\bar{X}$ we know that $\bar{X} = \Sigma fX/N$
i.e., $\bar{X} = 111/45$
So the actual mean of the observation is $\bar{X} = 2.47$

Step II, calculate the standard deviation by using the formula

$$SD\ (\sigma) = \sqrt{\frac{\Sigma fx^2}{N}}$$
$$= \sqrt{\frac{79.46}{45}}$$
$$= \sqrt{1.76}$$

So the deviation of weight among the forty five fishes is, $\sigma = 1.328$.

C. Continuous series

Example 4: Calculate the Standard deviation for the marks obtained by the students in their exam.

Marks	4-8	8-12	12-16	16-20
Frequency	2	5	8	3

Work procedure:

- Calculate the actual mean of the observations.
- Find the midpoint from the class.
- Calculate the deviation (x) of the value from the mean $(X-\bar{X})$.
- Square the deviation and calculate the Σx^2
- Multiple them by the respective frequencies and make the total i.e., Σfx^2

- Calculate the Standard deviation using the following formula.

$$SD\ (\sigma) = \sqrt{\frac{\Sigma f x^2}{N}}$$

Marks	frequency	mid point (m)	fm	$x = (m-\bar{X})$	x^2	fx^2
4-8	2	6	12	-6.67	44.48	89.17
8-12	5	10	50	-2.67	7.12	35.64
12-16	8	14	112	1.33	1.76	14.15
16-20	3	18	54	5.33	28.40	85.22
	N = 18		Σfm = 228			$\Sigma fx^2 = 224.18$

$\bar{X} = \Sigma fX/N$

$= 228/18$ or $= 12.67$

According to formula

$$SD\ (\sigma) = \sqrt{\frac{\Sigma f x^2}{N}}$$

$$= \sqrt{\frac{\Sigma 224.18}{18}}$$

$$= \sqrt{12.45}$$

Thus, $\sigma = 3.52$

Merits of Standard Deviation

1. Standard deviation rigidly defined and helps to summarises the deviation of a large distribution.
2. It is one of the most reliable measures of dispersion and used to observe the variations occur among the collected data.
3. The higher the value of standard of variation more the data have odd reading or more the data is fluctuating, so more the data in non-reliable in case of laboratory results.
4. It helps in finding the suitable size of sample for valid conclusions.

Demerits of Standard Deviation

1. Standard deviation includes very length mathematical calculations.
2. Standard deviation gives weightage to extreme values

Short Questions

1. Write a short note on Biostatistics.
2. Write difference between mean, median and mode.
3. Write merits and demerits of standard deviation.

4. Write short note on range and its utility.
5. Discuss the role of biostatistics in life sciences.
6. Discus various mean of dispersal.
7. Discuss measures of central tendency
8. Write short note on merits and utility of median.
9. Write the procedure for calculating assumed mean.
10. Name various measure of dispersal with their working formulas.
11. Write short note on Deciles.

6.3 Summary

The standard deviation is a statistical technique that's used to measure how much a group of data values vary or are dispersed. It displays the degree to which the mean (average) of the data set is different from each individual data point. A low standard deviation indicates a propensity of the data points to be near the mean, whereas a high standard deviation indicates a dispersion of the data points over a wider range. The standard deviation is frequently used in many different disciplines to evaluate the consistency and variability of data. Understanding the distribution and dependability of data is aided by this fundamental statistical idea.

6.4 Keywords

- Standard Deviation
- Actual Mean Method
- Assume Mean Method

6.5 Self-Assessment Questions

1. What is the definition of standard deviation?
2. How does standard deviation help in understanding data variability?
3. What is the difference between a low and high standard deviation?
4. What is the actual mean method in statistics?
5. How do you compute the mean of a dataset using the actual mean method?
6. What are the advantages of using the assumed mean method?
7. How do you calculate the standard deviation using the assumed mean method?

6.6 Case Study

Imagine a company analyzing the performance of its sales team over the last year. The monthly sales figures (in thousands) for each team member are recorded. The company wants to determine the consistency of sales performance.

- Sales figures for Team Member A: [50, 55, 52, 60, 48, 54, 53, 56, 51, 59]
- Sales figures for Team Member B: [40, 60, 70, 30, 80, 20, 50, 70, 40, 90]

Calculate the standard deviation for both Team Members A and B. Discuss what the standard deviation reveals about each team member's sales consistency.

6.7 References

1. Text Book of Biostatistics I. (2005). India: Discovery Publishing House Pvt. Limited.
2. Forthofer, R. N., Lee, E. S. (2014). Introduction to Biostatistics: A Guide to Design, Analysis, and Discovery. United States: Elsevier Science.

UNIT - 7

Coefficient of Variation

Learning Objectives

- Understand the Coefficient of Variation
- Understand the Interpretation
- Understand the Calculation method

Structure

7.1 Definition and interpretation

The coefficient of variation (CV) is a measurement of relative inconsistency often used to compare the dispersion of data across different datasets, especially when the datasets have different units or scales. It is computed as the percentage difference between a dataset's standard deviation (σ) and mean (μ).

$CV=\sigma\mu\times100\%$

Where:

- CV = Coefficient of Variation
- σ = Standard Deviation
- μ = Mean

The coefficient of variation provides a standardized measure of dispersion that is independent of the scale of the data, making it useful for comparing the variability of datasets with different units or scales. A higher CV shows larger relative inconsistency, while a lesser CV shows lower relative inconsistency.For example, if you have two datasets, A and B, with means of 50 and 100, and standard deviations of 10 and 20 respectively, the CVs would be calculated as follows:

For dataset A: CV_A=1050×100%=20%

For dataset B: CV_B=20100×100%=20%

In this example, both datasets have the same coefficient of variation, indicating that they have the same relative variability despite having different means and standard deviations.

7.2 Calculation method

Example 1: In a gymnosperms exam the marks obtained by the ten students is given, calculate mean deviation and coefficient from the data.

Marks	62	68	66	79	87	75	60	89	77	83

Arrange the data accordingly into the table and then calculate the arithmetic mean.

Calculate the mean deviation using formula, M.D. $=\frac{\Sigma\,|D|}{N}$

Then calculate the coefficient of mean deviation using the following formula,

Coefficient of M.D $=\frac{M.D}{Mean}$

Marks	Deviation from Mean (ignoring signs) |D |
62	12.6
68	6.6
66	8.6
79	4.4
87	12.4
75	0.4
60	14.6
89	14.4
77	2.4
83	8.4
ΣX = 746	Σ |D | = 84.8

Step I: Calculate the arithmetic mean Here we have,

ΣX = 746 and N = 10
Arithmetic mean = $\bar{X}$= ΣX/N
$\bar{X}$ = 746/10
= 74.6.

Step II: Calculate the deviation, ignoring signs
Calculate deviation is Σ |D | = 84.8

Step III: calculate mean deviation
M.D. = $\frac{\Sigma |D|}{N}$
M.D = 84.8/10, = 8.48
Thus, M.D is 8.48.

Step IV: Calculate coefficient of Mean deviation
Coefficient of M.D = $\frac{M.D}{Mean}$
= 8.48/74.6
= 0.113
In conclusion we have, Arithmetic mean = 74.6; Mean deviation = 8.48 and Coefficient of mean deviation from the data = 0.113.

7.3 Summary

In summary, the Coefficient of Variation is a powerful tool for comparing the degree of variation between different datasets, especially when the data have different units or widely different means. It provides a normalized measure of dispersion, making it easier to assess relative variability and consistency.

7.4 Keywords

- Coefficient of Variation
- Interpretation

- Mean
- Standard deviation

7.5 Self-Assessment Questions

1. What is the Coefficient of Variation (CV)?
2. How is the CV calculated? Provide the formula and explain each component.
3. Why is the CV considered a unitless measure?
4. What does a higher CV indicate about a dataset?
5. What does a lower CV suggest about the variability of data in relation to the mean?
6. In which scenarios would the CV be an inappropriate measure to use?

7.6 Case Study

You are a financial analyst tasked with comparing the risk and return profiles of three different mutual funds. The funds have the following annual return statistics:

Fund A:

- Mean return: 8%
- Standard deviation: 2%

Fund B:

- Mean return: 12%
- Standard deviation: 4%

Fund C:

- Mean return: 15%
- Standard deviation: 9%

Calculate the CV for Each Fund:

- Compute the CV for Fund A, Fund B, and Fund C.
- Show your calculations.

7.7 References

1. Text Book of Biostatistics I. (2005). India: Discovery Publishing House Pvt. Limited.
2. Forthofer, R. N., Lee, E. S. (2014). Introduction to Biostatistics: A Guide to Design, Analysis, and Discovery. United States: Elsevier Science.

UNIT - 8

Measures of Central Tendency for Qualitative Variables

Learning Objectives

- Understand the Central Tendency
- Understand the Qualitative data
- Understand Selection of appropriate methods of data collection

Structure

8.1 Mode for qualitative data

The most common value or values in a dataset are described by the mode, which is a measure of central tendency. When the values indicate groups or categories rather than numerical values, it is especially helpful for qualitative or categorical data.

To find the mode for qualitative data:

1. **Identify the Categories**: First, identify all the unique categories or values present in the dataset.
2. **Count the Frequency**: For each unique category, count how many times it occurs in the dataset.
3. **Find the Category with the Highest Frequency**: The category with the highest frequency is the mode or modes of the dataset. If there is only one category with the highest frequency, the data set is unimodal. If there are multiple categories tied for the highest frequency, the dataset is multimodal.

For example, consider a dataset representing the favorite colors of a group of people:

- Red
- Blue
- Green
- Blue
- Red
- Red

In this dataset, "Red" occurs three times, "Blue" occurs twice, and "Green" occurs once. Therefore, "Red" is the mode of the dataset.

It's important to note that unlike measures of central tendency such as the mean and median, the mode is not affected by outliers or extreme values since it only considers the frequency of categories. However, a dataset can have one mode, more than one mode , or no mode if all categories occur with equal frequency.

8.2 Other measures for categorical data

In addition to the mode, several other measures can be used to summarize and analyze categorical data:

1. **Frequency Distribution**: A frequency distribution table summarizes the number of times each category appears in the dataset. It provides a clear picture of the distribution of categorical data.
2. **Relative Frequency**: The ratio of the number of observations to the frequency of a category is known as relative frequency. One way to compute it is to divide the total number of observations in the dataset by the frequency of each category.
3. **Percentages**: Percentages are often used alongside frequencies to express the relative frequency of each category as a percentage of the total.
4. **Bar Charts**: The height of each bar in a bar chart, which represents categorical data graphically, corresponds to the frequency or relative frequency of each category. The frequencies of several categories can be visually compared with bar charts.
5. **Pie Charts**: Pie charts consist of circular graphs that have been split into slices. Each slice represents a category, and the size of the slice indicates how frequently or relatively frequently that category occurs. Pie charts are helpful in showing the relative amounts of various dataset groups.
6. **Measures of Association**: Measures of association, such as contingency tables and chi-square tests, are used to analyze relationships between categorical variables. They help determine whether there is a statistically significant association between two or more categorical variables.
7. **Cross-Tabulation (Contingency Tables)**: Cross-tabulation is a tabular method used to summarize the relationship between two categorical variables. It displays the frequency counts or percentages of observations for each grouping of categories of the two variables.
8. **Measures of Agreement**: Measures such as Cohen's Kappa are used to assess the agreement or reliability between two raters or methods when categorizing observations into different categories.

   These measures provide valuable acumens into the distribution, relationships, and agreement of categorical data, permitting researchers to make well-versedresolutions and draw meaningful deductions from their analyses.

8.3 Summary

- Mode is the primary measure of central tendency for qualitative data, identifying the most common category.
- Frequency Distribution and Proportion are complementary measures that provide insights into the relative occurrence and significance of each category.
- Median can be used when there is a natural or meaningful order to the categories, though it is less common.
- These measures help summarize and interpret qualitative data, offering a clear view of the most typical or popular categories within a dataset.

8.4 Keywords

- Qualitative data
- Measures of Agreement
- Cross-Tabulation
- Measures of Association

8.5 Self-Assessment questions

1. What are qualitative variables?
2. What are the primary measures of central tendency for qualitative variables?
3. What insights can be gained from the mode of a qualitative dataset?
4. How are proportions or percentages calculated for qualitative variables?
5. Provide an example of a frequency distribution for a qualitative variable.

8.6 Case Study

A company conducted a survey to understand customer preferences for four different smartphone brands. The results are as follows:

- Apple: 120 votes
- Samsung: 90 votes
- Google: 45 votes
- OnePlus: 45 votes

1. Determine the Mode:
    - Identify the mode of the dataset.
    - Explain why the mode is the appropriate measure of central tendency for this data.
2. Create a Frequency Distribution:
    - Construct a frequency distribution table for the survey results.
    - Calculate the proportion and percentage for each smartphone brand.
3. Interpret the Results:
    - Based on the mode and the frequency distribution, which smartphone brand is the most preferred among customers?
    - Discuss the significance of the proportions and percentages for each brand.
4. Customer Insights:
    - Provide insights and potential recommendations for the company based on the survey results.
    - How might the company use this information to inform their marketing strategy?

8.7 References

1. Text Book of Biostatistics I. (2005). India: Discovery Publishing House Pvt. Limited.
2. Forthofer, R. N., Lee, E. S. (2014). Introduction to Biostatistics: A Guide to Design, Analysis, and Discovery. United States: Elsevier Science.

UNIT - 9

Introduction to Karl Pearson's Coefficients of Correlation

Learning Objectives

- Understand the Karl Pearson's Coefficients
- Understand the Karl Pearson's Correlation

Structure

9.1 Karl Pearson's Coefficients

Karl Pearson's coefficients, which are commonly abbreviated as Pearson's coefficients, are a collection of statistical metrics that are employed to quantify the strength and direction of the association between two continuous variables. With a symbol of r, the Pearson correlation coefficient is the most often used Pearson's coefficient.

The Pearson correlation coefficient (r) indicates how strongly and linearly two variables are related to one another. The values it accepts range from -1 to 1:

If r = 1, it directs a perfect positive linear relationship, meaning that as one variable increases, the other variable also increases in a linear fashion.

If r = -1, it directs a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases in a linear fashion.

If r = 0, it directs no linear relationship between the variables.

Pearson's coefficient is calculated using the following formula:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where:

x_i and y_i are the individual data points.

$\bar{x}$ and $\bar{y}$ are the means of x and y respectively.

In order to evaluate correlations between variables, such as the association between height and weight, wealth and education level, or temperature and ice cream sales, Pearson's coefficient is extensively used in a variety of domains, including psychology, economics, biology, and social sciences.

It's important to remember that the normal distribution of the data and a linear relationship between the variables are presumptions made by Pearson's correlation coefficient. In cases when the data contradicts these presumptions, it might not be suitable. Furthermore, it should be noted that a large correlation between two variables does not indicate causality, which means that changes in one variable may not always result in changes in the other.

9.2 Correlation

In biostatistics, correlation refers to a statistical measure that describes the strength and direction of a relationship between two variables. Understanding correlation is crucial for identifying and quantifying the degree to which two variables are related. Here's a detailed look at correlation:

Type of correlation

1. Positive and Negative correlation
2. Simple and Multiple correlation
3. Partial and Total Correlation
4. Linear and Non Linear correlation

Positive and Negative Correlation:

The movement of variables in a single direction is referred to as positive correlation. It implies that as one variable increases, the other must likewise increase or decrease if one is declining. On the other hand, a negative correlation denotes the opposite direction of variable movement. Stated otherwise, as the value of one variable rises, the value of another variable tends to fall, or the opposite occurs.

Examples for positive correlation

a. Increase in heights and weight of a group of persons is a positive correlation.
b. The longer your hair grows the more shampoo you needs.
c. The more petrol you put in your bike, the farther it can go.
d. Price of commodity and amount of supply.
e. As the amount of moisture increases in an environment, the growth of mold spores increases.

Examples for negative correlation

a. The more one works, the less free time one has.
b. As a tadpole gets older, its tail gets smaller.
c. Demand of a commodity may go down as a result of rise in price.

Simple and Multiple Correlations: Simple correlation refers to the type of correlation in which only two variables are studied or in other words the relationship is confined to two variables. For example, when one studies relationship between the yield of rice and the area or land on which the seeds were sown. Multiple correlations, is a type of correlation in which there are more than two or in other words three or more variables are studied. For example, relationship of yield of rice, soil type, chemical and use of pesticides.

Partial Correlation: This type of correlation refers to the subtype of multiple correlation in which three or more variables but not all variables are consider. For example, the yield of rice is depend on nature of soil, water, fertilizer, type of seed and use of pesticides, but only two ormore variable used as rest are assumed to be constant.

Linear and Non-linear curvilinear correlation:

When the variation in the values of any two variables have a constant ratio, then the relationship is known as linear correlation, but if the values are not constant then the relationship is non-linear.

Correlation coefficient (r) ranges from -1 to 1:

- $r = 1$: Perfect positive correlation. The other variable grows in proportion to the growth in the first.
- $r = -1$: Perfect negative correlation. The other variable falls in direct proportion to the rise in the first.
- $r = 0$: No correlation. The two variables do not exhibit a regular linear connection.

Degree of correlation:

The degree or intensity of relationship between two variables can be classified into there

A. Prefect degree of correlation: This type of correlation shows the perfect relationship between two variables and with the fluctuation in the value of one the value of other is also changes.

B. Limited degree of correlation: In this type of correlation the variable shows low level of interdependence among each other or in other words there is unequal change in the

same or opposite direction.

C. Absence of correlation: In this type of correlation, there is no relationship exists between variables or in other words there is no interdependence between the two variables. Thus, this it indicate that there is no correlation or zero correlation between two variables.

Scatter diagram:

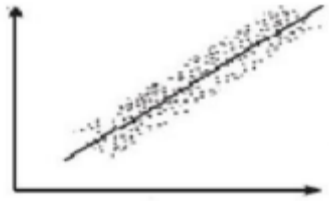

Figure 9.1 : Positive correlation

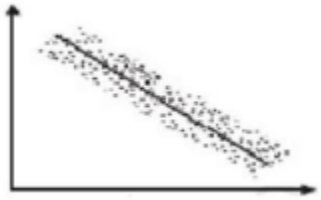

Figure 9.2 : Negative correlation

Figure 9.3 : No correlation

Figure 9.4 :Perfect positive correlation

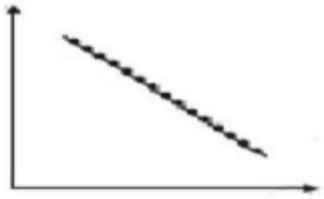

Figure 9.5 : Perfect Negative correlation

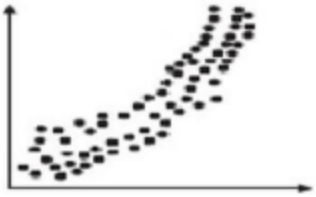

Figure 9.6 : Positive non-linear correlation

Example 1: Calculate coefficient of correlation from the given data and interpret the results.

Marks in Botany	34	22	16	23	24	21	15	19	26	30
Marks in Zoology	12	23	29	20	17	16	18	16	22	27

Solution:

1. Calculate the arithmetic mean of X and Y series.
2. Find out the deviation of X and Y.
3. Square these deviation and obtained Σx^2 and Σy^2 respectively.
4. Multiply the calculated deviation and find out the total Σxy.
5. Calculate coefficient using the given formula.

Marks in Botany	$X-\bar{X}$	x^2	Marks in Zoology	$Y-\bar{Y}$	y^2	xy
34	11	121	12	-9	81	99
22	-1	1	23	2	4	2
16	-7	49	29	8	64	56
23	0	0	20	-1	1	0
24	1	1	17	-4	16	4
21	-2	4	16	-5	25	10
15	-8	64	18	-3	9	24
19	-4	16	16	-5	25	20
26	3	9	22	1	1	3
30	7	49	27	6	36	42
$\Sigma X= 230$		**$\Sigma x^2= 314$**	**$\Sigma Y= 200$**		**$\Sigma y^2= 262$**	**$\Sigma xy= 260$**

According to formula:

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

Arithmetic mean of X and Y series can be calculated as

$\bar{X}=\frac{\Sigma X}{N}$ = 230/10 = 23.

$\bar{Y}=\frac{\Sigma Y}{N}$ = 200/10 = 20.

Further we have, $\Sigma xy= 260$, $\Sigma x^2= 314$ and $\Sigma y^2= 262$

$$r = \frac{260}{\sqrt{314 \times 262}}$$

$$r = \frac{260}{286.82}$$

Thus, r= 0.91

Hence, there is high degree of positive correlation.

Rank Correlation:

Even though we want to determine the significance of the relationship between the two variables, there are times when data are known not to follow the bivariate normal distribution. Calculating a coefficient of rank correlation and ranking the variates is one way to analyze such data. Rank correlation, sometimes referred to as Spearman correlation and represented by the symbol "r or ‘ (rho)", is a technique employed when quantifying characteristics such as natural beauty, color, stress, leadership potential, and personal knowledge becomes challenging. And was developed by British psychologist Charles Edward Spearman in 1904. In this method results cannot be measure quantitatively but ranks are allotted to each element either in ascending or descending order. Thus the developed formula of Spearman help in obtaining the correlation coefficient between ranks of n individual allotted in two series of ranks.

Formula, $$\mathbf{r \text{ or } \rho \text{ (rho)} = 1 - \frac{6\Sigma D^2}{n(n^2-1)}}$$

Where r or ρ (rho) = rank difference between of X and Y variables.

D= distinction between the two traits of the same person in a pair.

n= figure of pairs.

Example 2: Calculate the coefficient by rank method of the following data:

Marks in Botany	24	19	27	36	30	25
Marks in Zoology	37	29	28	31	33	24

Solution

S. No	Marks in Botany	Rank (R_1)	Marks in Zoology	Rank (R_2)	$d=R_1-R_2$	d^2
1	24	5	37	1	4	16
2	19	6	29	4	2	4
3	27	3	28	5	-2	4
4	25	4	31	3	-1	1
5	30	2	33	2	0	0
6	2536	1	24	6	-5	25
						$\Sigma d^2 = 50$

Thus according to formula: $r \text{ or } p \text{ (rho)} = 1 - \frac{6\Sigma D^2}{n(n^2-1)}$

$= 1 - \frac{6\times 50}{6(36-1)}$

$= 1 - \frac{300}{210}$

$= 1\text{-}1.43 = \text{-}0.43$

Since the obtained results showed negative rank correlation (-0.43), this indicates that the student who is best in one subject is worst in the other and vice-versa.

Uses of correlations

1. Correlation helps to study or draw a result in bivariate data.
2. Correlation analysis helps in deriving precisely the degree and the direction of relationship when two or more variable are involved.
3. Correlation helps in determining the individual difference and finding error in the perditions.
4. The measure of coefficient of correlation is a relative measure of change.
5. Ecological data and other bivariate data can be analysis by this method.
6. Correlation is important in many areas of measurement and evaluation on education.
7. Correlation help in determining Reliability in given data or condition.

9.3 Summary

Understanding these correlation concepts enables researchers and analysts to explore relationships between variables, identify patterns, and make informed decisions in various fields such as science, economics, and social sciences.

9.4 Keywords

- Correlation
- Karl Pearson’s Coefficients

9.5 Self-Assessment questions

1. What is Karl Pearson's correlation coefficient?
2. How is Pearson's coefficient calculated?
3. What does the value of Pearson's coefficient signify in terms of the relationship between variables?
4. Define positive and negative correlation.
5. How does the sign of Pearson's coefficient relate to positive and negative correlation?
6. Provide examples of scenarios demonstrating each type of correlation.

9.6 Case Study

A market inquiry firm leads a study to discover the relationship between advertising expenditure and product sales for a new line of smartphones. They collected data over six months and calculated Pearson's correlation coefficient.

Tasks:

1. Data Analysis:
    - Compute Pearson's correlation coefficient using the collected data on advertising expenditure and product sales.
2. Interpretation:
    - Interpret the value of Pearson's coefficient in terms of the relationship between advertising expenditure and product sales.
    - Discuss whether the relationship is positive, negative, or neutral.

9.7 References

1. Text Book of Biostatistics I. (2005). India: Discovery Publishing House Pvt. Limited.
2. Forthofer, R. N., Lee, E. S. (2014). Introduction to Biostatistics: A Guide to Design, Analysis, and Discovery. United States: Elsevier Science.

UNIT - 10

Concepts of Regression

Learning Objectives

- Understand the Simple linear regression
- Understand the Multiple linear regression
- Understand Interpretation of regression

Structure

10.1 Simple linear regression

In general data collection or sample survey the researcher is asked to relate two or more variables to predict an outcome. For example, in phyto-sociological samplings, how the vegetation varies with altitude for a given area. An example is the association between type of soil, vegetation, and forest type. Regression analysis and correlation are the names of the statistical techniques used to characterize or define these connections. Generally speaking, regression analysis is used to mathematically characterize a relationship between two variables, and correlation analysis offers a quantitative means of assessing the strength of that relationship. The ultimate objective is the creation of an equation for the prediction of one variable from one or more other variables. Regression analysis is a method, introduced by Fransis Galton for estimating or practicing the unknown value of one variable from known value of another. Regression analysis is a tool used to describe the connection between two or more variables, just like correlation analysis. In correlation analysis, correlation shows the degree and direction relationship between two variables, it does not clearly specify as to one variable is the cause and the other effect. "In regression analysis, the relationship between two variables (or more) is expressed by fitting a line or curve to the pairs of data points". In other words if we take a simple case of regression analysis we have to consider one dependent variable and one independent variable. For example, land holding of a family is related to the crop production for the family, so if we assume that land holding of a family increases the crop production for the family is also increases. From this we may say that, crop production is dependent variable and land holding is the independent variable. If we are going to denote the same on graph, we denote the dependent variable as Y and the independent variable as X.

Linear regression:

The link between a dependent variable and one or more independent variables may be modeled statistically using linear regression. To forecast results, comprehend correlations, and make defensible judgments, it is extensively utilized in several domains, including biostatistics.

Key Concepts of Linear Regression

1. **Dependent Variable (Y)**: The outcome or variable that you are trying to predict or explain.

2. **Independent Variable (X)**: The variable(s) that you use to predict or explain the dependent variable.
3. **Linear Relationship**: Linear regression assumes a linear relationship between the independent and dependent variables, which can be represented as a straight line.

Simple Linear Regression

In simple linear regression, we model the relationship between two variables using a linear equation:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Where:

- Y is the dependent variable.
- X is the independent variable.
- β_0 is the intercept of the regression line.
- β_1 is the slope of the regression line.
- ϵ is the error term, representing the deviation of the observed values from the predicted values.

Steps in Performing Linear Regression

1. **Data Collection**: Gather data on the dependent and independent variables.
2. **Model Fitting**: Use statistical software to fit the linear regression model to the data.
3. **Parameter Estimation**: Estimate the coefficients $(\beta_0, \beta_1, \ldots, \beta_n)$ that minimize the sum of the squared differences between the observed and predicted values of Y.
4. **Model Evaluation**: Assess the goodness of fit of the model using metrics like R-squared, adjusted R-squared, and the p-values of the coefficients.
5. **Prediction**: Use the fitted model to make predictions on new data.

Example Calculation

Let's consider an example dataset to illustrate simple linear regression:

Dataset:

- Independent Variable (X): [1, 2, 3, 4, 5]
- Dependent Variable (Y): [2, 3, 5, 7, 11]

We want to model the relationship between X and Y.

10.2 Multiple linear regressions

An expansion of simple linear regression, multiple linear regression models the connection between a dependent variable and two or more independent variables. This method is helpful for understanding how several factors affect an outcome and for predicting anything based on several variables in biostatistics and other related subjects.

Key Concepts of Multiple Linear Regressions

1. **Dependent Variable (Y)**: The outcome variable that you are trying to predict or explain.
2. **Independent Variables**$(X_1, X_2, \ldots, X_n)$: The variables that you use to predict or explain the dependent variable.
3. **Linear Relationship**: Multiple linear regressions assume a linear relationship between the dependent variable and each independent variable.

The Multiple Linear Regression Model

The model for multiple linear regressions is represented as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$$

Where:

- Y is the dependent variable.
- $X_1, X_2, \ldots, X_n$ are the independent variables.
- β_0 is the intercept.
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients for the independent variables.
- ϵ is the error term.

Steps in Performing Multiple Linear Regressions

1. **Data Collection**: Gather data on the dependent and independent variables.

2. **Model Fitting**: Use statistical software to fit the multiple linear regression model to the data.
3. **Parameter Estimation**: Estimate the coefficients $(\beta_0, \beta_1, \ldots, \beta_n)$ that minimize the sum of the squared differences between the observed and predicted values of Y.
4. **Model Evaluation**: Assess the goodness of fit of the model using metrics like R-squared, adjusted R-squared, and the p-values of the coefficients.
5. **Prediction**: Use the fitted model to make predictions on new data.

Let's consider an example dataset to illustrate multiple linear regression:

Example 1:

Dataset:

- Independent Variables (X1: Hours Studied, X2: Hours Slept)
- Dependent Variable (Y: Test Scores)

Hours Studied (X1)	Hours Slept (X2)	Test Scores (Y)
10	6	85
9	7	80
8	5	78
7	8	70
6	5	65

We want to model the relationship between the hours studied and slept (independent variables) and the test scores (dependent variable).

10.3 Interpretation of regression results

Interpreting regression results involves understanding the relationship between variables, assessing the significance and magnitude of coefficients, and evaluating the overall fit and validity of the model. Here are the key steps in interpreting regression results:

1. **Coefficients**:

- **Magnitude**: To comprehend how independent factors affect the dependent variable, look at the coefficient magnitudes. A higher coefficient denotes a more robust correlation.
- **Significance**: Assess the significance of each coefficient using hypothesis tests, typically with t-tests or F-tests. A significant coefficient suggests that the independent variable has a non-zero effect on the dependent variable.

2. **Interpretation of Coefficients**:
   - In a basic linear regression, the dependent variable changes for each unit change in the independent variable, as shown by the coefficient, with other variables being maintained constant.
   - Each coefficient in multiple linear regression represents the change in the dependent variable for every unit change in the corresponding independent variable, while keeping all other variables constant.

Standard Errors and Confidence Intervals:

- Standard errors indicate the precision of coefficient estimates. Smaller standard errors imply greater precision.
- Confidence intervals give a range of values that the real population parameter should fall inside. A greater confidence interval implies more uncertainty.

3. **Goodness-of-Fit**:
   - ***R*2 (Coefficient of Determination)**: Designates the proportion of discrepancy in the dependent variable enlightened by the independent variables. Higher $R2R2$ values suggest a well fit.
   - **Adjusted *R*2**: Adjusts for the number of predictors in the model. It penalizes excessive use of predictors and provides a more conservative estimate of goodness-of-fit.
   - **F-statistic**: Assesses the overall significance of the regression model. A significant F-statistic indicates that at least one independent variable has a significant effect on the dependent variable.
4. **Residual Analysis**:
   - Evaluate the residuals (the differences between observed and predicted values).

- Check for patterns in residuals using residual plots to ensure that assumptions such as linearity, normality, and constant variance (homoscedasticity) are met.
- Look for outliers or influential data points that may unduly influence the results.

5. **Model Comparison**:
   - Compare different models using criteria such as $R2R2$, adjusted $R2R2$, and model complexity (number of predictors).
   - Consider parsimony and interpretability when selecting the best-fitting model.

Example 2: Calculate the two regression for X and Y or Y on X for the following data.

X	1	4	6	8	10	3	5	7	9
Y	10	12	14	16	18	24	20	15	13

Solution

X	**Y**	**X^2**	**Y^2**	**XY**
1	10	1	100	10
4	12	16	144	48
6	14	36	196	84
8	16	64	256	128
10	18	100	324	180
3	24	9	596	72
6	20	36	400	100
7	17	49	289	119
9	13	81	169	117
$\Sigma X = 54$	**$\Sigma Y = 144$**	**$\Sigma X^2 = 392$**	**$\Sigma Y^2 = 2474$**	**$\Sigma XY = 878$**

Here, n=9, for both X and Y

In first case: Regression equation of X on Y: **X = a + bY**

The two normal equations are:

$\Sigma X = Na + b\Sigma Y$

$\Sigma XY = a\Sigma Y + b\Sigma Y^2$

Substituting the values in above normal equations, we get

54= 9a+ 144b ……..(i)

878= 144a+ 2474b ……...(ii)

Let us solve these equations (i) and (ii) by simultaneous equation method

Multiply equation (i) by 16 we get 864 = 144a + 2304b

Now rewriting these equations:

864= 144a+ 2304b

878= 144a+ 2474b

(-) = (-) + (-)

-14 = -170

Therefore now we have -14 = -170b this can rewritten as 170b = 14

Now, b =14/170 = 0.082(round off)

Substituting the values of b in equation (1), we get

54 = 9a+144(0.082)

9a = 54 – 11.85

a = -4.68

Now the regression X and Y is

X = -4.77+0.082Y

In Second case: Regression equation for Y on X

Y = a +bx

The two normal equations are:

$\Sigma Y = Na + b \Sigma X$

$\Sigma XY = a \Sigma X + b \Sigma X^2$

Substituting the values in above normal equations, we get

144 = 9a+54b ……...(1)

878=54a+392b …….(2)

Let us solve this equation (1) and (2)

$$864 = 54a + 324b$$

$$878 = 54a + 392b$$

$$(-) = (-) + (-)$$

$$-8 = -68$$

Therefore now we have -8 = -68b, this can rewritten as 68b = 8

Now b =8/68 =0.117

Substituting the values of b in (1), we get

144 =9a+54(0.117)

144 =9a+6.318

a = 15.29

Therefore a = 15.29

So the two equations are

X = -4.77+0.082Y

Y = 15.29+0.117X

10.4 Summary

1. Interpretation of Regression:
   - Focuses on understanding how changes in independent variables impact the dependent variable.
2. Simple Linear Regression:
   - Predicts a dependent variable using one independent variable.
   - Interpretation involves understanding the effect of the independent variable on the dependent variable.
3. Multiple Linear Regression:
   - Predicts a dependent variable using two or more independent variables.
   - Interpretation involves understanding the effects of multiple independent variables on the dependent variable while holding other variables constant.

10.5 Keywords

- Interpretation of regression
- Multiple linear regression
- Simple linear regression

10.6 Self-Assessment questions

1. What is the primary goal of interpreting regression analysis results?
2. Why is it essential to understand the coefficients, significance levels, and goodness-of-fit measures in regression analysis?
3. Define simple linear regression and its key components.
4. How is the coefficient interpreted in simple linear regression?
5. Provide an example scenario where simple linear regression would be appropriate.
6. Define multiple linear regression and its key components.
7. How are the coefficients interpreted in multiple linear regression?
8. Provide an example scenario where multiple linear regression would be more appropriate than simple linear regression.
9. How can regression analysis be applied in real-world scenarios, such as business, healthcare, or social sciences?

10.7 Case Study

A retail company is interested in predicting monthly sales constructed on numerous factors such as advertising expenditure and store location. They have collected historical data on sales, advertising spending, and location attributes for the past year.

Tasks:

1. Simple Linear Regression:
    - Conduct a simple linear regression analysis to predict monthly sales based on advertising expenditure.
    - Interpret the coefficients and assess the significance of the regression model.
2. Multiple Linear Regression:

- Conduct a multiple linear regression analysis to predict monthly sales based on advertising expenditure and store location attributes.
- Interpret the coefficients and assess the significance of the regression model.

10.8 References

1. Text Book of Biostatistics I. (2005). India: Discovery Publishing House Pvt. Limited.
2. Forthofer, R. N., Lee, E. S. (2014). Introduction to Biostatistics: A Guide to Design, Analysis, and Discovery. United States: Elsevier Science.