Bachelor of Commerce
(B.Com.)

BUSINESS STATISTICS
(DBCMCO302T24)

Self-Learning Material
(SEM III)

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

Prof. R.L. Raina
Former Vice Chancellor, JKLU
Professor of Communication, IIM Lucknow

Prof. J.K. Tandon
Former Dean, Faculty of Commerce
Former Head of EAFM Department
University of Rajasthan

Course Coordinator

Ms. Namrita Singh Ahluwalia
Assistant Professor
Department of Business & Management, JNU, Jaipur

Unit Preparation

Unit Writers

Ms. Pinky Arora
Department of Business & Management, JNU, Jaipur
(Unit 1-4)

Dr. Harish Kumar
Department of Business & Management, JNU, Jaipur
(Unit 5-8)

Dr. Vicky Likhar
Department of Business & Management, JNU, Jaipur
(Unit 9-12)

Assisting & Proof Reading

Dr. Abhishek Raizada
Department of Business & Management, JNU, Jaipur

Editor

Prof. Prashant Madan
Jaipur National University, Jaipur

Secretarial Assistance

Mr. Nitin Parashar

Course Introduction

Business Statistics is assigned 5 credits and contains 12 units. Application of Business Statistics makes the manager to apply statistical methods to analyze and interpret business data effectively

The decisions taken on the basis of Business Statistics are subject to evaluation and objective assessment.

Each unit is divided into sections and sub-sections. Each unit beginswith statement of objectives to indicate what we expect you to achieve through the unit.

Course Outcomes

After studying this course, a student will be able to:

1. Identify the key terminology, tools and techniques used in business statistical analysis and able to recall statistical concepts.
2. Demonstrate the underlying usage of Central Tendency of data dispersion of data.
3. Determine the uses and limitations of Correlation and Regression
4. Relate critically summarize and illustrate Index Numbers
5. Appraise the problems using the techniques covered and conduct basic Statistical Analysis of Time Series and Association of Attributes.
6. Formulate the decision-making power related to estimates about cost, demand, prices, sales etc.

We hope you will enjoy the course.

Acknowledgement

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

Statistical Data and Descriptive Statistics

CHAPTER – 1: Introduction to Data and statistics

Learning Objective:

After Studying the unit, Students will be able to:

- Know what statistics is.
- Understand the meaning of statistics and data.
- Explain the types of data sources.

Structure

1.1 Concept

1.2 Definition of stats

1.3 Types

1.4 Functions

1.5 Summary

1.6 Self-Assessment Questions

1.7 References/ References Reading

1.1 Introduction

"Statistics" defines as numerical data that has been expressed quantitatively. It's possible for this information to be about things, people, events, phenomena, or geographical areas. In actuality, there are no restrictions on the extent, coverage, or references of data. These statistics provide information on the gross domestic product (GDP) and the respective proportions of manufacturing, agriculture, and services in the GDP at a macroeconomic level. Regardless of their size, individual firms generate significant statistical data regarding their activities at a micro level. Various statistics on sales, output, expenses, inventories, capital utilized, and other activities are included in annual reports of firms.

These data are frequently field data that have been gathered using scholarly survey methods. These facts are the outcome of a singular endeavor and have restricted utility beyond the specific conditions that prompted their gathering, unless they are regularly refreshed. Statistics, as a discipline akin to economics, mathematics, chemistry, physics, and other fields, allows students to develop a more profound comprehension of the subject. It is a discipline that works with data in a scientific manner and is frequently referred to as the science of data. As a result of dealing with statistics as data, a body of these methods for gathering, presenting, summarizing, and analyzing data has been established in statistics.

1.2. DEFINITIONS AND MEANING OF STATISTICS

It should be noticed at the outset that the word "statistics" is employed fairly oddly in both plural and singular senses. It refers to a group of numbers or information when used in the plural. Statistics encompass a comprehensive set of tools utilized to collect, arrange, and analyze data, with the ultimate goal of deriving meaningful insights from it. It is important to emphasize that in order for quantitative data to be valuable, both sides of statistics are crucial. If statistics were an inadequate discipline with ineffective methods, we would be unable to ascertain the optimal approach for extracting information from the data. Likewise, even if our subject is very advanced, if our data is faulty, insufficient, or inaccurate, we will be unable to reach accurate conclusions. According to A.L. Bowley, “statistics are:

(i) Statistics is the science of counting,

(ii) Statistics may rightly be called the science of averages, and

(iii) Statistics is the science of measurement of social organisms regarded as a whole in all its mani- festations.”

According to Boddington, “Statistics is the science of probabilities and estimations.”“The science of statistics, as described by W.I. King in a broader sense, is the process of evaluating

collective, natural, or social phenomena based on the outcomes of an analysis, enumeration, or collection of estimates." According to Seligman's research, "statistics is a science that deals with the techniques for gathering, categorizing, presenting, contrasting, and interpreting numerical data in order to shed light on any area of study." According to Spiegal, "statistics is the scientific method for gathering, organizing, summarizing, presenting, and analyzing data as well as for drawing accurate conclusions and making sensible decisions based on such analysis."'Spiegel emphasizes the role statistics plays in decision-making, particularly under uncertainty.''Prof. Horace Secrist' defined "statistics as the collection of facts that are significantly impacted by a variety of causes, numerically expressed, counted, or estimated in accordance with reasonable standards of accuracy, systematically gathered for a predetermined purpose, and arranged in relation to one another."

Followingabove definitions provided, we may emphasize the main traits of statistics:

(i) Statistics are collections of factual information. It indicates that a solitary number does not qualify as statistical data. For instance, the country's national income during one year does not qualify as statistical data, whereas the national income for two or more consecutive years does.

(ii) Various factors influence statistical outcomes. The sale of a product is determined by various factors, such as its price, quality, level of competition, consumer income, and other variables.

(iii) Statistics need to be substantially accurate. If incorrect numbers are analyzed, incorrect conclusions will result. As a result, conclusions must be supported by reliable data.

(iv) Statistics must be gathered in a methodical way. Haphazard data collection will result in unreliable data that will support false findings.

(v) Collected methodically for a pre-established objective. Statistics should also be compared or evaluated in relation to each other. It is unclear and impossible to draw any logical conclusions from data that have no relation to one another. Data should be comparable across geography and time.

1.3. TYPES

Data is the fundamental component. Data refers to a specific occurrence, a problematic situation being studied, or an action that captures our attention. They arise from the processes of measurement, enumeration, and/or observation. Statistical data refers to the measurable, quantifiable, tallyable, or categorizable parts of a problem scenario. An object subject

variable refers to any phenomena or activity that produces data. In essence, a variable is an entity that exhibits fluctuating levels of alteration across multiple observations. Statistics categorizes data into two main classifications: quantitative data and qualitative data.

Quantitative data refers to information that may be measured using particular units of measurement. These are characteristics that can be quantified multiple times to get observable measurements. Quantitative data can be classified as either continuous or discrete, depending on the inherent characteristics of the quantity being measured.

Undoubtedly, a variable can be classified as either continuous or discrete.

(i) Continuous data display a continuous variable's numerical values. A continuous variable represents an interval of values, allowing for any value to be taken between two locations on a line segment. Despite their high level of accuracy and strong correlation, the values are evidently distinct from each other. All measurable qualitiesare classified as continuous variables. Therefore, data that is recorded based on these specific attributes and comparable features is referred to as continuous data. It is important to prioritise the use of an ideal unit of measurement for a variable that can take on any value within a range. The phrase "finest" denotes the capacity of a measurement to attain the highest degree of precision.

(ii) Discrete data is the value that a discrete factor assumes. Any resulted variablewhich iscalculated in discrete numbers is referred to as discrete. Such information is just a count. These are collected by means of a quantitative method, like counting the quantity of products that possess or lack a particular attribute. Discrete data examples include the daily traffic at a department store, the arrivals at an airport, and the defective goods in consignments that have been accepted for sale.

Qualitative traits of a subject or an object are referred to as qualitative data. When an observation is characterized and reported in terms of the presence or absence of a specific property in discrete numbers, it is said to be qualitative in nature. This information is further divided into nominal and rank information.

(i) Nominal data are derived from the process of categorising the elements or entities within a sample or population into two or more distinct groups, based on certain criteria related to their quality. Nominal data are produced by categorizing workers by skill (as skilled, semi-skilled, and unskilled), students by gender (as men and females), and employees by educational level (as matriculants, undergraduates, and

post-graduates). Any such basis for classification allows for the easy assignment of each object to a specific class and the tally of items from each class. Nominal data are the count data that were obtained in this way.

(ii) In other way round, Rank data is a method used to assign a numerical order to a set of values, often represented by integers 1, 2, 3,..., n. Ranks may be assigned based on the level of achievement achieved on an exam. A competition, an interview, a show, or a contest. There are two categories of data sources: secondary and main. The two fit the following definitions:

(i) Secondary data: They are data that have already been obtained from a secondary source and are either published or unpublished in some manner. They can usually be obtained from publicly available sources;however they may not be available in the exact format required.

(ii) Primary data: Information that must be gathered for the first time from a primary source because it does not currently exist in any form (s). These data must be brand new and collected for the first time, encompassing the entire population or a sample taken from it.

1.4. FUNCTIONS OF STATISTICS

The following are some of statistics' usage:

1. It demonstrates verifiable information in a precise manner: Given that mathematical expressions are compelling, one of the fundamental responsibilities of statistics is to depict factual data in a straightforward and succinct manner.
2. It facilitates the understanding of large quantities of data: Information presented in the form of a table, graph, diagram, average, or coefficients is readily comprehensible.
3. It facilitates comparison: After simplifying the data, it may be compared to other comparable data. These comparisons would not have been feasible without the data.
4. It aids in forecasting: Organizational plans and policies are typically developed in advance prior to their execution. Gaining insight into future trends is essential for formulating efficient policies and plans.
5. Statistical procedures such as z-test, t-test, and x2-test are highly valuable for establishing and testing hypotheses, as well as for developing new theories.
6. Statistics provide essential data for the development of acceptable policies. Given the possibility of alterations, it facilitates the estimation of export, import, or manufacturing initiatives.

7. Statistics indicate trend behavior: Time series analysis, regression, correlation, and other statistical approaches are helpful in predicting the future.

1.5 Summary

"Statistics" refers to numerical data that has been expressed quantitatively. It's possible for this information to be about things, people, events, phenomena, or geographical areas. In actuality, there are no restrictions on the extent, coverage, or references of data.

It should be noticed at the outset that the word "statistics" is employed fairly oddly in both plural and singular senses. It refers to a group of numbers or information when used in the plural. Statistics encompasses the complete range of tools utilised for collecting data, arranging and examining it, and subsequently deriving conclusions from it as a whole. It should be focused that for the quantitative data to be useful, both aspects of statistics are critical. If statistics were an inadequate discipline with ineffective methods, we would be unable to determine the optimal approach for extracting information from the data. Likewise, even if our subject is very advanced, if our data is faulty, insufficient, or inaccurate, we will be unable to derive accurate conclusions.

According to Boddington, “statistics is the science of probabilities and estimations.”“The science of statistics, as described by W.I. King in a broader sense, is the process of evaluating collective, natural, or social phenomena based on the outcomes of an analysis, enumeration, or collection of estimates.” According to Seligman's research, “statistics is a science that deals with the techniques for gathering, categorizing, presenting, contrasting, and interpreting numerical data in order to shed light on any area of study.”

The fundamental building block of statistics is data. Data may relate to a phenomenon, a problem scenario under research, or an activity of our interest. They result from the measurement, counting, and/or observational processes. Therefore, the elements of a problem scenario that can be measured, quantified, tallied, or categorized are referred to as statistical data. An object subject variable is any phenomenon or activity that generates data through this procedure. Put simply, a variable is something that shows varying levels of change throughout multiple observations.

1.6 Self-Assessment Questions

1. Define the nature of data?
2. Write the 4 types of data classification?
3. What do you mean by classification data?

4. What is the nature of data in statistics?
5. Why is data classification important?
6. What are the types of data?
7. What is the meaning of statistics?
8. What is data in statistics?
9. What is the definition of data?
10. Differentiate between bivariate and multivariate data?

1.7 References / Reference Reading

1. Lakshmikantham, D.; Kannan, V. (2002). Handbook of stochastic analysis and applications. New York: M. Dekker.

2. Schervish, Mark J. (1995). Theory of statistics (Corr. 2nd print. ed.). New York: Springer.

CHAPTER - 2: Classification of Data

Learning Objective:

After Studying the unit, Students will be able to:

- Know what different data analysis entails.
- Understand the meaning of time series and cross sectional data.
- Explain the different types of data.

Structure

2.1. UNIVARIATE ANALYSIS

CLASSIFICATION OF DATA

There are three primary analysis strategies that can be used in statistics, which vary based on the level of analysis required. These include

- “Univariate analysis”
- “Bivariate analysis”
- “Multivariate analysis”

. The next section is an explanation of the three distinct phases of data analysis.

Univariate analysis is the predominant method employed in statistical data analysis.
Univariate analysis is used when there is just one variable in the data and there are no causal links or effects.
Here's an illustration of a univariate analysis:

The researcher may be attempting to enumerate the quantity of male and female students in a classroom while performing a survey. Under these circumstances, the data would solely depict a solitary variable and its magnitude, as illustrated in the table provided below. The primary objective of Univariate analysis is to succinctly elucidate the data and identify inherent patterns. This can be achieved by analysing several statistical measures such as the mean, median, mode, dispersion, variance, range, standard deviation, and so on.
There are various methods for conducting univariate analysis, most of which are descriptive in nature.

- Distribution of FrequencyTabulation
- Histographical Representation
- Polygonic Representation of Frequency
- Pie Representation
- Bar Representation

2.2.BivariateAnalysis

Bivariate analysis is marginally more analytical. Bivariate analysis is the appropriate analytical approach when a dataset contains two variables and researchers intend to compare the two datasets.

Below is a straightforward illustration of bivariate analysis. –

The researcher's objective in conducting a classroom survey is to examine the distribution of kids who achieved scores above 85%, based on their gender. In this context, gender is treated as an independent variable denoted by X, while the outcome is represented by Y as the dependent variable.

(Bivariate analysis) is performed by utilizing-
Coefficients of Correlation
Analysis of Regression

2.3. MULTIVARIATE ANALYSIS

When the data set contains more than two variables, a more intricate statistical analysis technique known as multivariate analysis is employed.
An illustration of multivariate analysis is shown here.
An individual in the medical profession has gathered information regarding the levels of cholesterol, blood pressure, and weight. Furthermore, she gathered data regarding the individuals' dietary patterns, including their weekly consumption of food. To comprehend how each variable in this situation relates to the others, a multivariate analysis would be necessary.

Typical multivariate analysis methods include:

- "Factor Analysis"
- "Cluster Analysis"
- "Variance Analysis"
- "Discriminant Analysis"
- "Multidimensional Scaling"
- "Principal Component Analysis"
- "Redundancy Analysis"

2.4. TIME- SERIES AND CROSS-SECTIONAL DATA

Time series data can be studied to determine the chronological progression of variables, distinguishing it from other types of data. Time is a critical aspect since it not only shows the variations in data among data points, but also has an impact on the outcomes. It offers an additional data source and a predetermined sequence of data interdependencies.

Time series analysis often necessitates a significant volume of data to assure consistency and reliability. Ensuring a comprehensive data collection is crucial for efficiently filtering out conflicting data and correctly representing the population in your study. Moreover, it ensures that any existing patterns or trends are not abnormal and can accommodate for seasonal variations. It is used for the purpose of forecasting, which entails creating predictions about future events by analysing past information.

WHY ORGANIZATIONS USE TIME SERIES DATA ANALYSIS

Time series analysis enables organizations to gain a deeper understanding of systemic patterns over time. Business users have the ability to analyse seasonal patterns and gain a deeper understanding of their underlying causes by utilising data visualisations. Today's analytics solutions offer visualisations that go beyond simple line graphs, providing a wide range of capabilities.

Organisations can utilise time series forecasting to assess the probability of future occurrences by analysing data at consistent intervals. Predictive analytics encompasses the forecasting of time series data. It can identify anticipated alterations in data, such as seasonal or cyclical patterns, so enhancing forecasting accuracy and providing a deeper comprehension of data variables.

Des Moines Public Schools conducted an analysis of student accomplishment data spanning five years in order to identify kids who are in a vulnerable position and monitor their progress. Thanks to the advancements in technology, it has become increasingly convenient to collect a sufficient amount of dependable data for comprehensive analysis. Currently, we have the capability to gather vast quantities of data on a daily basis.

TIME SERIES ANALYSIS EXAMPLES

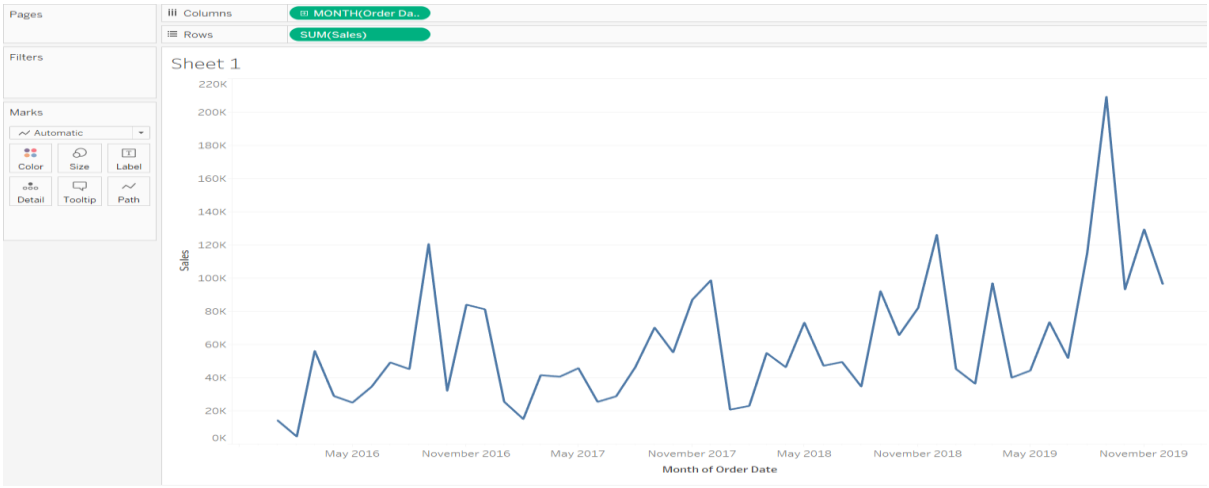


Fig.: 2.1 "Analysis of Time Series"

Time series analysis is used to study non-stationary data, which refers to data that changes over time or is influenced by time. Time series analysis is frequently employed in fields such as banking, retail, and economics due to the constant fluctuations in currency and sales. Utilising automated trading algorithms provides a remarkable demonstration of time series analysis in action within the context of stock market analysis. Time series analysis is highly effective in forecasting. Here are some examples of time series analysis being applied:

- Weather Forecasting
- Rainfall Measurement
- Measurements of temperature
- Electrocardiogram (EKG) for monitoring heart rate
- Electroencephalography (EEG) for monitoring brain activity
- Sales on a quarterly basis
- Share pricing

TIME SERIES ANALYSIS TYPES

It encompasses diverse data categories and variances, necessitating analysts to occasionally develop complex models. Analysts are unable to make generalizations about a certain model that apply to all samples or account for all variations. A lack of fit can arise when models are excessively complex or seek to accomplish multiple tasks simultaneously. Inadequately fitted or overfitted models are unable to distinguish between genuine relationships and random error, distorting analysis and generating erroneous predictions.

Time series analysis models encompass a variety of techniques and methodologies.

- Classification: The process of determining and assigning specific categories to the data.
- Curve fitting is a method used to analyse the relationships between variables in a dataset by charting the data points along a curve.
- Descriptive analysis involves the identification of trends, cycles, or seasonal variations in time series data.
- Explanatory analysis: Engages in a diligent endeavour to comprehend the data, its interconnections, and the correlations of cause and effect.

Exploratory analysis: Utilises exploratory analysis techniques to identify and emphasise the fundamental characteristics of the time series data, often shown visually.

- Forecasting: Anticipates future events by utilising forecasting techniques. This type is constructed based on previous patterns. It predicts possible results at various plot points by utilising historical data as a template for future data.
- Intervention analysis examines how a scenario can modify the data. Segmentation is a process that involves breaking data into smaller parts in order to uncover the fundamental attributes.

DATA CLASSIFICATION

Furthermore, it can be categorized into two distinct groups:

- Stock time series data represents the specific attributes of a stock at a particular moment, much like a photograph of the data at that exact time.
- The user's text is a bullet point. The term "flow time series data" is used to describe the measurement of attribute activity over a specific time period, typically reported as a percentage of the results.

DATA VARIATIONS

Irregular variations can occur in time series data.

- Functional analysis is a method that can be used to identify significant events by examining patterns and connections in the data.
- Trend analysis involves identifying a consistent pattern in a specific direction. There are two distinct categories of trends: stochastic, characterised by randomness and lack of rationality, and deterministic, where we can discern the underlying reason.

- Seasonal variation refers to events that occur regularly and at predetermined periods throughout the year. Serial dependence arises when data items that are near in time are likely to be linked.

In time series analysis and forecasting models, it is necessary to clearly describe the exact types of data needed to address the business question. Analysts select the appropriate types of analysis and methodologies to employ once they have chosen the pertinent data they wish to evaluate.

IMPORTANT CONSIDERATIONS FOR TIME SERIES ANALYSIS

Itmeans the data that is collected in specific time interval, and it may be identified through different types of data that reflect the specific time and method of acquisition. As an example: A time series refers to data that has been collected consistently over a fixed period of time.

- Multiple factors are concurrently documented in cross-sectional data.
- 'Pooled data' encompasses both cross-sectional and time series data.

TIME SERIES ANALYSIS MODELS AND TECHNIQUES

There are multiple methodologies available for analysing data, just as there are various categories and frameworks. Here are the three highest-ranking items:

Univariate Box-Jenkins analysis ARIMA models are utilised for forecasting future data points of variables and comprehending a solitary time-dependent variable, such as temperature throughout time. In order for these models to be effective, it is crucial that the data exhibits stationarity. Analysts should meticulously assess and eliminate as many variations and seasonal trends from historical data sets as feasible. The ARIMA model integrates moving averages, seasonal difference operators, and autoregressive components. Box-Jenkins models, which are multivariate models, are used to assess several time-dependent variables throughout time, such as temperature and humidity.

The Holt-Winters method utilises exponential smoothing as its strategy. If there is a detectable pattern of recurring fluctuations in the data points, the objective is to forecast or anticipate the outcomes.

CROSS-SECTIONAL DATA

The cross-sectional study utilises data collected at a single point in time to examine a specific population or phenomenon. Cross-sectional data refers to information collected by simultaneously examining various entities such as organisations, governments, regions, and individuals. An approach to analyse cross-sectional data involves examining the disparities among the individuals or groups being studied.

Cross-sectional data refers to information that is gathered from all participants simultaneously. However, it is important to note that in a cross-sectional study, the participants do not all provide their replies simultaneously.

Participants' cross-sectional data is collected more efficiently. This time period is alternatively referred to as the "field period." The passage of time does not alter the outcomes, but it does cause them to diverge.

By incorporating daily sales revenue and expenses into your data gathering process for several months, you will obtain a time series representing sales and expenses.

CROSS SECTIONAL DATA ILLUSTRATION

Consider a single scenario. If you desire to ascertain the current level of individuals' blood pressure, you may proceed. A random selection of 1000 individuals will be made from that group. It is alternatively referred to as a cross section of that particular demographic. Next, we will evaluate the individual's blood pressure. In addition, their height, weight, and other relevanthealth information will be documented.

This dataset offers a thorough perspective of the entire population. This data will solely offer the current prevalence of blood pressure values. It is not possible to accurately assess if the rate of blood pressure increase is low or excessive based on a single cross-sectional sample.

However, it will provide you with a comprehensive understanding of the forthcoming events.

Another instance of cross-sectional data is a study that examines the many ice cream flavours offered at a certain store and analyses people's responses to those flavours. A cross-sectional analysis can be derived from a compilation of individual student performance on a certain test within a class.

Data was collected regarding the sales, sales volume, expenses, and customer count of a coffee shop for the previous month. This is another form of cross-sectional data by including daily sales revenue and expenses into your data gathering process.

Comparing Data Types

Fig: 2.2 Various Data Types

Data can be of different shapes and sizes. This information measures various things at diverse points in time. Well, financial analysts are especially interested in both time-series data and cross-sectional data.

Various types of data are analysed using distinct methodologies. It is crucial to possess the ability to distinguish between time series and cross-sectional datasets. Let's talk about each one in turn and figure out what makes them different.

CROSS SECTIONAL DATA

These are the simultaneous observations made by several organisations or people. The individuals of the underlying population should possess comparable characteristics. For instance, if you wish to ascertain the quantity of organisations that commit resources towards research and development.

Many companies use less money on research and development than others. This will give you different information because different companies are in different groups. Instead, you can look at the companies in the same group and do a cross-sectional analysis on them. Time-series data is what we'll talk about next.

TIME-SERIES DATA

These are recurring observations conducted systematically at fixed intervals.

Time series data can be visualized with a graph that displays the weekly sales of ice cream in a business during the summer. Another example is the number of staff at a college, which was recorded every month. It was done to find out how often people quit their jobs. In the near future, these instancesmay be used to show how data patterns can be seen.

Let's try to make it clearer. Time-series data is a type of data that is gathered for the same variable over time, like months or years. The data could be collected over months or years, but they can be seen for almost any length of time.

Get online statistics assignment help if you need more help with cross-sectional and time-series data.

USES OF CROSS-SECTIONAL DATA

Differential equations and statistical methods use cross-sectional data. Cross-sectional regression is the main way it is used. It's similar to a regression analysis, but with this data. For example, a person's costs for using things in a certain period may be regressed based on various factors.

These factors may encompass their earnings, their assets, and their various demographic characteristics. This is to find out how differences among these characteristics affect how people act in the end.

Some practical examples of cross-sectional data

- Cross-sectional data is predominantly utilised in finance, economics, and various other domains within the social sciences.
- Cross-sectional data are utilised in the field of applied microeconomics.
- Political academics utilise cross-sectional data to analyse and dissect demographic characteristics and voter engagement.
- When comparing the financial statements of two or more companies, cross-sectional data can also help. This is what financial analysts do for a living. In time series data analysis, on the other hand, a company's financial statements from different times are compared.
- Cross-sectional data is very important in the retail business. It can look at how men and women of any age spend their money over time.
- Cross-sectional data can be used in business to look at how people from different socioeconomic backgrounds in a certain area respond to a single change.

- Cross-sectional data can also be used in medicine and health care to figure out how many kids between the ages of 4 and 14 are likely to have too little calcium.

THE CONCEPT BEHIND ROLLING CROSS-SECTION

In a rolling cross-section, both the fact that a person is in a sample and the interval during which he was in that sample are determined by random methods. After the selection, each person is given a date chosen at random. It is a random piece of information that is used to decide who to interview, so it is a part of the survey.

MERITS

- It takes less time to do a cross-sectional data study.
- All the information for this study is gathered at the same time.
- At the same time, research can be done on more than one outcome.
- It's a good way to gather information for descriptive analysis.
- It can help you start or continue your research.

DEMERITS

- At times, it could be hard to find the people who have the same factors.
- Associations are hard to figure out.
- The study can also have a bias.
- It doesn't help to figure out why.

2.5 Summary

Qualitative traits of a subject or an object are referred to as qualitative data. When an observation is characterized and reported in terms of the presence or absence of a specific property in discrete numbers, it is said to be qualitative in nature. This information is further divided into nominal and rank information.

Univariate analysis is the primary technique used in statistical data analysis. Univariate analysis is employed when there is a single variable in the data and no causal or consequential relationships are present.

When the data set contains more than two variables, a more intricate statistical analysis technique known as multivariate analysis is employed.

2.6 Self-Assessment Questions

1. Define rolling cross-section?
2. What are the merits of rolling cross sections?
3. Write the demerits of rolling cross sections?
4. What are the uses of cross-sectional data?
5. Why is data classification important?
6. What are the types of data?
7. What is univariate data?
8. What is bivariate data?
9. Define multivariate data?
10. Differentiate between bivariate and multivariate data?

2.7. References

- Moses, Lincoln E. (1986) Think and Explain with Statistics, Addison-Wesley, ISBN 978-0-201-15619-5. pp. 1–3
- Hays, William Lee, (1973) Statistics for the Social Sciences, Holt, Rinehart and Winston, p.xii.
- Moore, David (1992). "Teaching Statistics as a Respectable Subject". In F. Gordon; S. Gordon (eds.). Statistics for the Twenty-First Century. Washington, DC: The Mathematical Association of America. pp. 14–25.

CHAPTER – 3: Measures of Central Tendency

Learning Objective:

After Studying the unit, Students will be able to:

- Know what measures of central tendency are.
- Understand the different means and their uses.
- Analyze using mean, median and mode.

Structure

3.1 Mode

It is sometimes referred to as a measure of centre or central placement, is a succinct statistical measure that seeks to describe an entire dataset by using a single number.

Central tendency measures

It is referred to as a measure of centre or central position, is a succinct summary statistic that seeks to condense an entire database into a single value that indicates the middle or centre of its distribution.

The mode, median, and mean are the basic approaches for selecting the most prominent value. Each of these metrics represents a unique understanding of the mean or prevailing value.

The mode represents the most often occurring value within a given dataset.
Here is a dataset displaying the retirement age of 11 individuals, measured in complete years:
54, 54, 55, 55, 56, 57, 55, 57, 60, 58, 58

Age	Frequency
54	2
55	3
56	1
57	2
58	2
60	1

The value 55 comes up most often, so 55 years is “mode” of this distribution.

ADVANTAGE OF THE MODE:

The mode surpasses both the median and the mean because to its applicability in both numerical and non-numerical datasets.

LIMITATIONS OF THE MODE:

There are limitations on the permissible usage of the mode. The mode may not be a reliable measure to determine the central tendency of a distribution in certain cases. When arranging

the retirement age in ascending order, it becomes evident that the median age is Fifty Seven years, but the mean is shorter at 55.

54, 54, 55, 55, 56, 57, 55, 57, 60, 58, 58

It is also be the possibility for the same set of data to have more than one mode (bi-modal, or multi-modal). The mode cannot describe the center or typical value of the distribution when mode is more than one because it can't find a single value that describes the center.

3.2. Median

It means the center value of distribution in increasing or decreasing order.

It is the middle point of the distribution. Half of the observations are on either side of the median. The median value is the value that is in the middle when there are an odd number of observations.

When we look at the distribution of retirement ages that has ELEVEN points, the median 57 years:

54, 54, 55, 55, 56, 57, 55, 57, 60, 58, 58

The median value is the average of the middle two values when the no. of observations is even:

54, 54, 55, 55, 56, 57, 55, 57, 60, 58, 58

ADVANTAGE OF THE MEDIAN:

The median is a preferable measure of central tendency compared to the mean when the distribution is not symmetrical. This is because the median is less susceptible to the influence of outliers and skewed data.

LIMITATION OF THE MEDIAN:

The median cannot be determined for categorical nominal data due to the absence of a logical ordering system.

3.3. Mean

The mean is extracted out by summing up all value in a set and then dividing the total by the total observations. This is often referred to as the mathematical mean.

Taking another look at the distribution of retirement ages;

54, 54, 55, 55, 56, 57, 60, 58, 58

ADVANTAGE OF THE MEAN:

The mean is applicable to both continuous and discontinuous quantities.

LIMITATIONS OF THE MEAN:

- With categorical data, you can't figure out the mean because you can't add up all the values.
- Since the mean takes into account every value in the distribution, outliers and skewed distributions can change it.

WHAT MORE INFORMATION SHOULD I BE AWARE OF ABOUT THE CONCEPT OF "THE MEAN"?

The Greek letter "mu" represents the mean number of individuals in a group. The mean of a sample distribution is represented by the symbol $\bar{x}$, pronounced as X-bar.

3.4 Skewed Distribution

The mode, median, and mean of a symmetrical distribution are all located at the centre of the distribution. The next graph displays an expanded range of retirement ages with a symmetrical distribution on both sides. The mode, the middle, and the average are all 58 years (Fig. 3.1, 3.2& 3.3).

Fig. 3.1. Skewed Distributions

In a skewed distribution, the mode retains its status as the value that occurs most frequently, the median remains the value located in the middle, but the mean is usually pulled towards the extremes. In a skewed distribution, the mean is usually not positioned at the centre of the distribution, making the median a more appropriate measure of central tendency.

A dissemination is deemed to be tilted rightif the length of the tail on the right end is longer than the length of the tail on the left side. In a distribution with positive skewness, the mean is typically biassed towards the right side of the distribution. Although there may be a few exceptional cases, most values, including the median, typically have a smaller magnitude in comparison to the mean.The following graph displays an expanded dataset on retirement age, with a distribution that is skewed towards the right. Given that the variable being assessed, which is the age of retirement, is continuous, the data has been categorized into classes. The mode age is 54, the mode age range is 54-56, the median age is 56, and the mean age is 57.2.

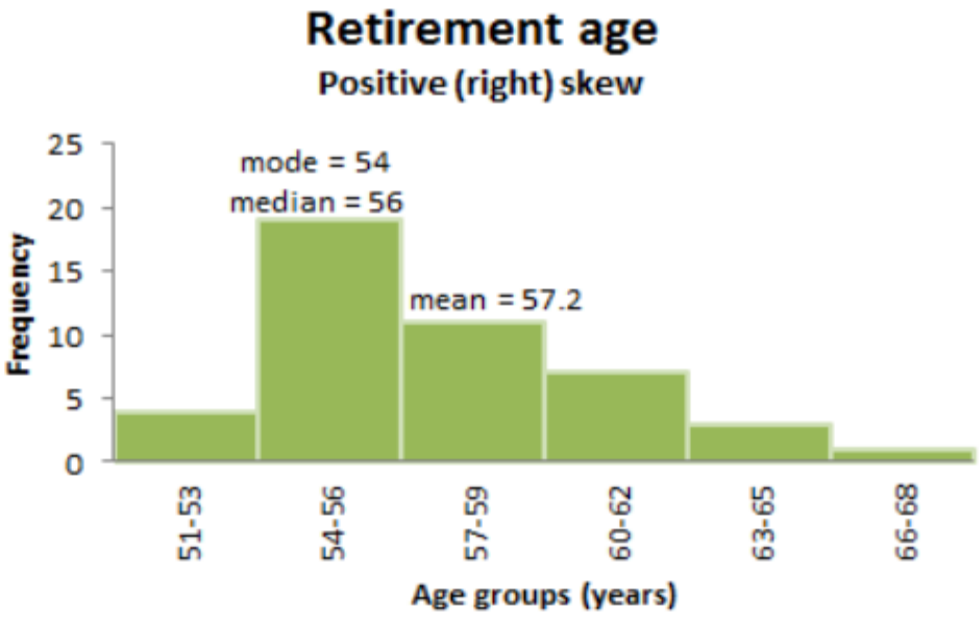


Fig. 3.2. Positive Skewed Distributions

If the left tail is higher than the right tail, the dissemination is said to be negatively skewed or left skewed. The mean of a negatively skewed distribution is pushed to the left side of the distribution. It's possible that some values won't follow this trend, but most of the time, values like the median are higher than the mean. The next graph shows a bigger set of data on retirement age, with a range that leans to the left. The age range with the most occurrences is 63–65, with 65 being the most common. The median age is 63, and the mean age is 61.8.

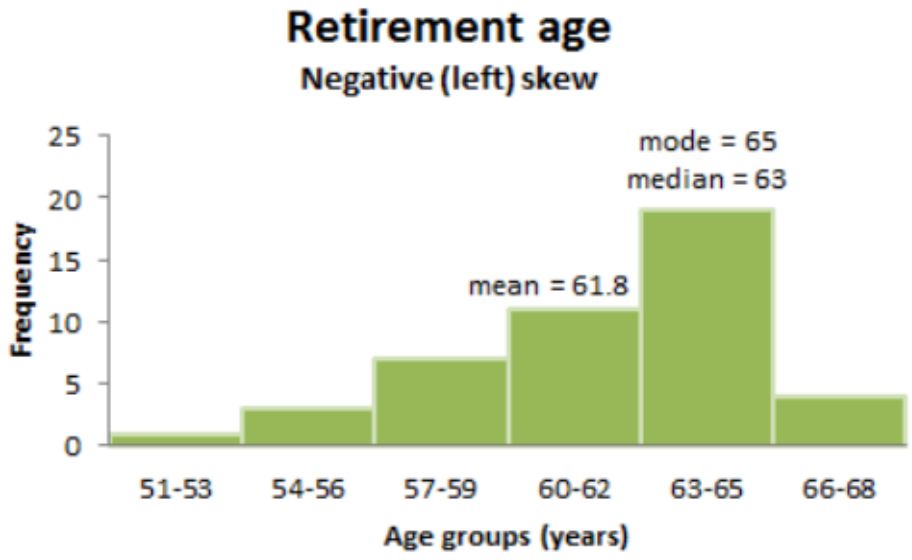


Fig. 3.3. Negative Skewed Distributions

FACTORS DEVIATING MEASURES OF CENTRAL TENDENCY

Outliers points are very different from the rest of the data.Finding outliers in a set of data is important.

Think about the first set of data on retirement age again, but this time the last observation of 58 years has been changed to 81 years. This value is a lot higher than the others, so it might be an outlier. The middle of the distribution hasn't changed, though, so 57 years is still the median value.

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 81

As each and every value is used to figure out the mean, the outlier will have an effect on the mean value.

(644 + 54 + 54 + 55 + 56 + 57 + 58 + 60 + 81) / 11 = 58.5 years

The presence of an outlier in this distribution has resulted in an increase in mean. The mean can still serve as a reliable indicator of the central tendency, particularly when the remaining data follows a normal distribution. It is verified to be a legitimate extreme number, it should not be excluded from the database. Various conventional methods can be employed to mitigate the impact of outliers on the average value.

TYPES OF MEANS

The three fundamental Pythagorean means consist of the arithmetic mean (AM), the geometric mean (GM), and the harmonic mean (HM).

The harmonic mean is the smallest among the geometric and arithmetic means. minimum

Arithmetic mean

It can be calculated by adding up all the numbers in the dataset and then divide by the total number of events.

As an illustration, the arithmetic mean is calculated by adding 4, 10, and 7, resulting in a sum of 21. Dividing this sum by 3 gives a mean of 7.

The math works well when the numbers are related in a way that adds them up. This frequently occurs when the numbers exhibit a "linear" relationship, indicating that when

plotted on a graph, they cluster along or in close proximity to a straight line. However, it is important to note that not all datasets exhibit a linear connection. Occasionally, one may anticipate a relationship that is either multiplicative or exponential in nature. Under such circumstances, the arithmetic mean is an inadequate method for summarizing the data and has the potential to be misleading.

GEOMETRIC MEAN

The geometric mean is effective when the data exhibits a multiplicative relationship or when the data is combined through addition. To restore the product to the dataset's range, multiplication is used instead of addition. The data serves as a multiplier and does not contain any missing or negative numbers. The geometric mean is employed when the data exhibits a non-linear pattern, such as when a logarithmic modification is applied to the data.

Suppose you invest $500 in an asset that yields a 10% return in the first year, and soon. After a period of three years, the amount of money you will have will be $858.00, calculated by multiplying $500 by 1.1, then by 1.2, and finally by 1.3.

By employing the arithmetic mean, the average annual return would amount to 20% based on the values of 10, 20, and 30. Consequently, after three years, the total amount would be calculated as $500 multiplied by 1.4 raised to the power of 3, resulting in $1372. It is evident that the arithmetic mean overestimates earnings by around $514, which is incorrect due to the inclusion of an addition operation in a procedure that was intended for multiplication. Typically, investors use the use of geometric mean rather than arithmetic mean to assess the performance of an investment or portfolio.

HARMONIC MEAN

Harmonic mean is used to find the average of things like speeds, rates, or ratios.

For example, I drove to downtown Seattle at a speed of 60 km/h and back home at a speed of 30 km/h. It is 20 miles from my house to downtown Seattle. How fast did I drive most of the time?

Harmonic Mean as stated

For us, A = 60 and B = 30. So, the "harmonic mean" is 40 km/h.

If you took the arithmetic mean of the two speeds, 45 km/h, that would not be right.

Therefore, to choose the correct mean for the correct process is important.

3.5. SUMMARY

When the distribution isn't symmetrical, the median is a better way to find the central trend than the mean. The middle doesn't get skewed or affected by outliers as much as other groups do.There is no logical order in categorical nominal data, so the middle cannot be found. The mode is the value that happens most often in a skewed distribution, the median is the value in the middle, but the mean tends to move towards the extremes. The median is a better way to find the central trend in a skewed distribution because the mean is not always in the middle of the distribution.

3.6. SELF-ASSESSMENT QUESTIONS

1. Mention the three types of bivariate analysis?
2. Why is bivariate regression used?
3. What is meant by time series data?
4. What is an example of time series data?
5. How do you analyze time series data?
6. What are the limitations of time series?
7. How do you describe a time series graph?
8. What is arithmetic mean, geometric mean and harmonic mean?
9. How do you find the geometric mean and arithmetic mean?
10. What is the relationship between arithmetic and harmonic sequence?

3.7. REFERENCE READINGS

- Anderson, D.R.; Sweeney, D.J.; Williams, T.A. (1994) Introduction to Statistics: Concepts and Applications, pp. 5–9. West Group.
- "Journal of Business & Economic Statistics". Journal of Business & Economic Statistics. Taylor & Francis. Archived from the original on 27 July 2020. Retrieved 16 March 2020

CASE STUDY

The questions in the worksheet pertain to the determination of the mean, median, and mode.

1. Calculate the average of the given data.

(a) 9, 7, 11, 12, 4, 5, 4, 4

(b) 18, 16, 19, 22, 21, 24, 27, 29, 29, 35

(c) 3.2, 11.2, 15.7, 5.9, 4.9, 10.1, 11.5

(d) 21/4, 31/2, 41/2, 31/4, 11/2

CHAPTER – 4: Measures of Variation

Learning Objective:

After Studying the unit, Students will be able to:

- Know what measures of variation are.
- Understand the Advantages and disadvantages of standard deviation.
- Explain the meaning and uses of various measures of deviation.

Structure

4.1 ABSOLUTE MEASURE OF DISPERSION

Measures of Variation are used in statistics to figure out how different the data is, or how homogeneous or heterogeneous it is. In plain English, it indicates how tight loose the variable could be.

MEASURES OF VARIATION

In statistics two main types of dispersion methods are there which are:

* Absolute Measure of Dispersion

It shows how far apart a set of observations are from each other. These numbers measure the spread of the data in the same units as the data itself. You can't use absolute measures to analyze the differences between multiple series or data sets. It doesn't tell you by itself if the difference is big or small.

Absolute Measures of Dispersion that are used a lot are:

- "Range"
- "Quartile Deviation"
- "Mean Deviation"
- "Variance or Standard Deviation"

These absolute measures of dispersion or spread are as follows:

"RANGE"

It means the distinction in between the biggest number in the data set and the smallest number. If the smallest value in a set of ungrouped data is Xo and the largest value is Nn, then the range (R) is defined as

Rn = Xn – Xo

For grouped data, there are three ways to figure out the range.

"R= Midpoint of the highest class – midpoint of the lowest class"

"R = Upper class limit of the highest class – Lower class limit of the lower class"

"R = upper class boundary of the highest class – Lower class boundary of the lowest class"

Arithmetic Mean and Range in Statistics

In statistics, the arithmetic mean is a common way to show how a group of numbers fits together. The arithmetic mean is sometimes called the average or just "mean."The mean is the number that is in the middle of a set of numbers. This gives us the arithmetic mean. When you add up all the observations, you get the arithmetic mean.

4.2 Quartile Deviation (Semi-Interquartile Range)

The distinction between the 3rd and 1st quartiles is called the Quartile deviation. 1/2 of this difference is called the semi-interquartile range (SIQD) or just the quartile deviation (QD).

The Quartile Deviation is better than the Range because it doesn't depend on how big or small the observations are. However, it doesn't tell you anything about the position of observations that fall outside of the two numbers. It can't be solved mathematically, and sampling variability has a big effect on it. Even though Quartile Deviation isn't used very often as a measure of dispersion, it is used when extreme observations are thought to be wrong or not representative. Since the Quartile Deviation is not based on all observations, extreme observations can change it.

One way to measure dispersion is with the quartile deviation. Before we go into more detail, let's review what quartiles are and how to define them. The values Q1, Q2, and Q3 are examples of quartiles. They take a list of numbers and split it into thirds. The middle part of the three quarters displays the data values that are close to the middle point and finds the median, which is the middle value in the distribution. Half of the data points are below the median, as shown by the lower part of the quarters. The other half are above the median, as shown by the upper part. In other words, the quartiles show how spread out the data set is.

In statistics, the term "quartile deviation" refers to the statistic that measures the spread. In this case, the Dispersion is the state of being spread out. Dispersion in statistics is a measure of how far numbers are likely to vary from an average value. In other words, dispersion helps you figure out how the data are spread out.

Quartile Deviation Definition

Mathematically, the difference between the upper quartile and the lower quartile is equal to half of the Quartile Deviation. Here, QD stands for quartile deviation.

QUARTILE DEVIATION FORMULA

Suppose Q1 is the lower quartile, Q2 is the median, and Q3 is the upper quartile for the given data set, then its quartile deviation can be calculated using the following formula.

QD = (Q3-Q1)/2

QUARTILE DEVIATION FOR UNGROUPED DATA

For an ungrouped data, quartiles can be obtained using the following formulas,

"Q1 = [(n+1)/4]th item"

"Q2 = [(n+1)/2]th item"

"Q3 [3(n+1)/4]th item"

Where n reflects the total observations in database.

Also, the median of the given data set is Q2, while the median of the lower half is Q1 and the median of the upper half is Q3.

Before we can figure out the quartiles, we have to put the given data values in order from lowest to highest. If n is an even number, we can use the same method to find the middle number.

Quartile Deviation for Grouped Data

"For a grouped data, we can find the quartiles using the formula,

$$\boldsymbol{Q_r = l_1 + \frac{r(N/4) - c}{f}\,(l_2 - l_1)}$$

Here,

Or the rth quartile

11 = the lower limit of the quartile class

12 = the upper limit of the quartile class

f = the frequency of the quartile class

c = the cumulative frequency of the class preceding the quartile class

N = Number of observations in the given data set"

4.3 Mean Deviation (Average Deviation)

"It is the arithmetic mean of the deviations from the mean or the median". All of these differences are counted as positive to avoid a problem caused by the fact that the sum of the differences between observations and the mean is zero.

“It can be found for other central tendencies, but when deviations are taken as the median, it is the easiest to figure out.”

“The Mean Deviation tells you more than the Range or the Quartile Deviation because it is based on all the values” that have been seen. The Mean Deviation doesn't give big differences too much weight, so it should probably be used when big differences are likely to happen.

“Step 1: Find the mean value for the given data values”
“Step 2: Now, subtract the mean value from each of the data values given”
“Step 3: Now, find the mean of those values obtained in step 2.”

Formula for Mean Deviation Here is the formula to figure out the mean deviation for the given set of data. Mean Deviation = $[\Sigma |X – \mu|]/N$

Here,

- “Σ represents the addition of values”
- “X represents each value in the data set”
- “μ represents the mean of the data set”
- “N represents the number of data values”
- “|| represents the absolute value, which ignores the "-" symbol”

“Mean Deviation for Frequency Distribution”

In order to present the data in a more discrete way, we put it into groups and say how often each group appears. Class intervals are what people call these groups.
There are two ways to put data into groups:
Discrete Frequency Distribution
Continuous Frequency Distribution
Let's start by figuring out what the discrete distribution of frequency actually means.

Mean Deviation for a Discrete Distribution Frequency

By "discrete," we mean that something is separate or doesn't go on forever. In this kind of distribution, “the number of observations (or frequency) given in the set of data is discrete.
If the data set consists of values x1,x2, x3...xn each occurring with a frequency of f1,

f2... fn respectively then such a representation of data is known as the discrete distribution of frequency."

The following steps are taken to figure out the mean deviation for grouped data, and especially for data with a discrete distribution:

Step I: The measure of central tendency about which mean deviation is to be found out is calculated. Let this measure be a.

If this measure is mean then it is calculated as,

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

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

Where,

$$N = \sum_{i=1}^{n} f_i$$

If the measure is median, the given set of data is put in ascending order and the cumulative frequency is calculated. The observations whose cumulative frequency is equal to or just above N/2 are taken as the median for the given discrete frequency distribution, and it can be seen that this value is in the middle of the frequency distribution.

Step II: Find the absolute difference between each observation and the measure of the central tendency from Step 1.

Step III: The formula is then used to figure out the mean absolute deviation from the measure of central tendency.

$$M.A.D(a) = \frac{\sum_{i=1}^{n} f_i |x_i - a|}{N}$$

If the central tendency is mean then,

$$M.A.D(\bar{x}) = \frac{\sum_{i=1}^{n} f_i |x_i - \bar{x}|}{N}$$

In case of median

$$M.A.D(M) = \frac{\sum_{i=1}^{n} f_i |x_i - M|}{N}$$

Mean Deviation Example

Determine the mean deviation for the data values 5, 3,7, 8, 4, 9.

Solution:

Given data values are 5, 3, 7, 8, 4, 9.

We know the procedure to calculate the mean deviation.

First, find the mean for the given data:

Mean, μ = (5+3+7+8+4+9)/6

μ = 36/6

μ = 6

Therefore, the mean value is 6.

Now, subtract each mean from the data value, and ignore the minus symbol if any

5-6 = 1

3-6 = 3

7-6 = 1

8-6 = 2

4-6 = 2

9-6-3

Now, the obtained data set is 1, 3, 1, 2, 2. 3.

Finally, find the mean value for the obtained data set

Therefore, the mean deviation is

=(1+3+1+2+2+3)/6

= 12/6

= 2

Hence, the mean deviation for 5, 3.7, 8, 4, 9 is 2.

Variance and Standard Deviation

In statistics, the two most important measurements are the variance and the deviation. Variance is a way to measure how far apart data points are from the mean.

Variance

In simple terms, the variance is a way to measure how far apart a set of data is from its mean or average value. It is written as ‘σ2’

It can never be negative because each term in the sum of the variances is squared, so the result can only be positive or zero.

The units of the variance are always squared. For instance, the difference between a set of estimated weights in kilograms is given as kg squared.

4.4 Standard Deviation

The standard deviation is a way to measure how spread out statistical data is. Distribution is a way to figure out how far data is from its mean or average. The method for estimating the difference between data points is used to figure out the degree of dispersion. It is shown with the symbol."

It is also called the root-mean-square deviation. It is the square root of the mean of the squares of all the values in a data seta

Since it can't be less than 0, 0 is the smallest value of the standard deviation.

Formula

The average square distance between the mean value and each data value is set variance. And the standard deviation shows how far apart the data points are from the mean.

	Population	Sample
Variance	$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$	$S^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$
Standard deviation	$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$	$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$

The formulas are given below:

Variance Formula:

The population variance formula is given by:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (X_i - \mu)^2$$

Here,

σ2= Population variance

N = Number of observations

Xi = ith observation

μ= Mean of Population

The sample variance formula is given as:

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

Here,

s2 = Sample variance

n = Number of observations in sample

xi = ith observation in the sample

x̄ = Sample mean

Standard Deviation Formula

The population standard deviation formula is given as:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (X_i - \mu)^2}$$

Here,

σ = Population standard deviation

Similarly, the sample standard deviation formula is:

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

Here,

s = Sample standard deviation

Properties of Standard Deviation

Itmeans the statistical measure of how scattered apart the values in a data collection are. It is calculated by taking the square root of the average of the squared differences between each value and the mean of the data set. The standard deviation is synonymous with the root mean square deviation. This is an alternative designation for it.
The only feasible value for the standard deviation that is less than zero is zero itself. This is due to the fact that the standard deviation cannot have a negative value.
When the data sets are comparable, the standard deviation values are either close to zero or have a small absolute value. However, if there is a substantial difference between the values of these data sets or if they have a considerable discrepancy, then the standard deviation will be substantially higher.

4.5. Summary

Measures of Variation in statistics are employed to determine the extent of dissimilarity or homogeneity/heterogeneity within the data. Put simply, it indicates the degree of variability of thevariable.

The Absolute Measure of Dispersion quantifies the extent of variation among a group of data. These statistics quantify the extent of dispersion in the data using the same units as the data itself.
In statistics, to find the range, we need to put the given values, data, or observations in ascending order. That means you should first write down the observations from least to most important.
The arithmetic mean is a common way to show how a group of numbers fits together. The arithmetic mean is sometimes called the average or just "mean."
Mathematically, the difference between the upper quartile and the lower quartile is equal to half of the Quartile Deviation. Here, QD stands for quartile deviation, where Q3 is the upper quartile and Q1 is the lower quartile.
To present the data in a more compact way, we put it into groups and say how often each group appears. Class intervals are what people call these groups.

In simple terms, the variance is a way to measure how far apart a set of data is from its mean or average value. It is written as ‘σ2’

The standard deviation is a way to measure how spread-out statistical data is. Distribution is a way to figure out how far data is from its mean or average. The method for estimating the difference between data points is used to figure out the degree of dispersion. It is shown with the symbol.

4.6. Self-Assessment Questions

1. Define the relation between mean deviation and standard deviation?
2. What is mean deviation?
3. What does deviation mean?
4. How do we calculate mean deviation?
5. What is the difference between mean deviation and standard deviation?
6. What is the formula of mean deviation from median?
7. What does the standard deviation mean?
8. Which is better, high or low standard deviation?
9. What is the Formula of Mean Deviation?
10. What does deviation mean for grouped data?

4.7 References

- McCarney R, Warner J, Iliffe S, van Haselen R, Griffin M, Fisher P (2007). "The Hawthorne Effect: a randomized, controlled trial". BMC Med Res Methodol. 7 (1): 30.
- Rothman, Kenneth J; Greenland, Sander; Lash, Timothy, eds. (2008). "7". Modern Epidemiology (3rd ed.). Lippincott Williams & Wilkins. p. 100.
- Mosteller, F.; Tukey, J.W (1977). Data analysis and regression. Boston: Addison-Wesley.

CASE STUDY

Given the following data, calculate mean, variance and standard deviation:

Class Interval	0-10	10-20	20-30	30-40	40-50	50-60
Frequency	27	10	7	5	4	2

UNIT III: SIMPLE CORRELATION AND REGRESSION ANALYSIS

CHAPTER 5 - Correlation Analysis

Learning Objective:

After studying the unit, students will be able to:

- Know what correlation is.
- Understand the meaning of correlation analysis.
- Explain the types of correlation.

Structure

5.1. Correlation Analysis

It is a statistical method used to measure the linear relationship between two variables and establish their connection. It quantifies the degree to which changes in one variable are impacted by changes in another one. A strong correlation between two variables indicates a robust link. A low correlation signifies a weak connection between two variables.

Market analysts utilize correlation analysis to analyses the quantitative data gathered from sources such as surveys and live polls. They examine patterns, significant correlations, and trends between two variables or sets of data.

A positive correlation refers to the phenomenon when an increase in one variable is accompanied by a corresponding increase in another variable. In contrast, a negative correlation refers to a situation where an increase in one variable is accompanied by a decline in the other variable, and vice versa.

5.2. Meaning of Correlation

Itmeans how closely two variables are linked in a straight line (meaning they change together at a constant rate). It reveals about simple connections without saying anything about cause and effect.

5.3 How is Correlation Measured?

The strength of the relationship is shown by the sample correlation coefficient, r. Correlations are also looked at to see if they are statistically important.

What are Some Limitations of Correlation Analysis?

Correlation can't look at the effects or presence of variables other than the two being studied. Most importantly, correlation doesn't tell us anything about what causes what.

Correlations describe data moving together

Correlations are a good way to describe simple connections between data. Say, for example, you are looking at a set of data about campsites in a mountain park. You want to know if the campsite's elevation (how high up the mountain it is) has anything to do with the average high temperature in the summer.

There are two ways to measure each campsite: altitude and temperature. When you use a correlation to compare these two things across your sample, you can find a linear relationship: as elevation goes up, temperature goes down. They go against each other.

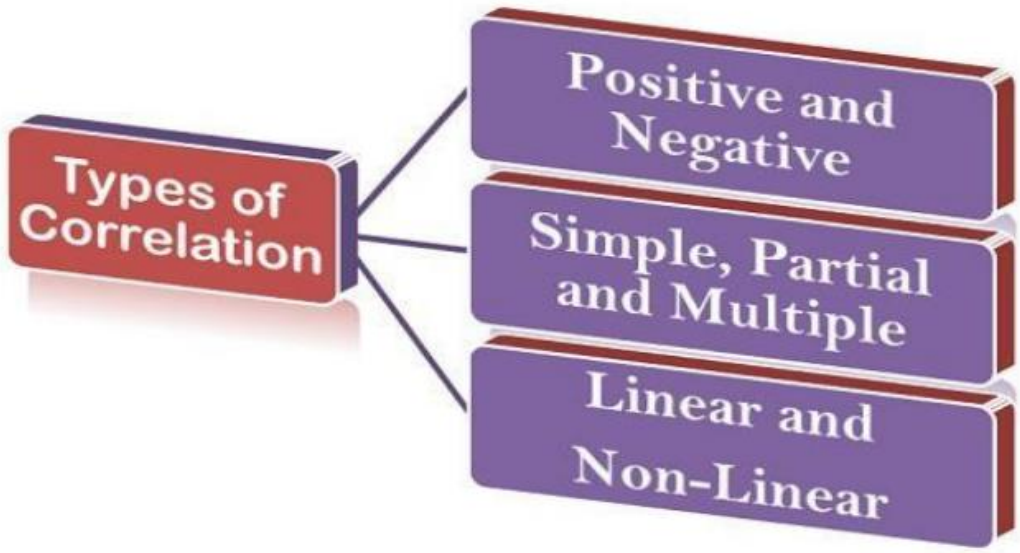
What Do Correlation Numbers Mean?

The correlation coefficient, that ranges from -1 to +1 and is written as r, is a measure that doesn't have any units. A p-value shows if something is statistically important. Because of this, correlations are usually written with two important numbers: r = and p =.

- The linear relationship gets less strong as r gets closer to 0.
- When the r value is positive, it means that the values of both variables tend to go up at the same time.
- When the value of one variable goes up when the value of the other variable goes down, this is called a negative correlation.
- Based on what we see in the sample, the p-value shows that it is likely that the correlation coefficient for the whole population is different from zero.

After identifying a meaningful link, we can further examine its level of strength. However, in reality, it is highly unlikely to observe a perfect correlation between variables unless one variable is a direct substitute for the other. Observing a correlation coefficient that is precisely perfect indicates the presence of an error in your data. For instance, if you recorded the altitude instead of the temperature for each campsite, it would precisely correspond to the elevation.

5.4 Types of Correlation



Simple, Partial and Multiple Correlations

How simple, partial, or many the correlation is depending on how many variables are studied. When only two variables are being looked at, the correlation is seen as simple. When three or more variables are being looked at, the correlation is either multiple or partial. When three variables are looked at the same time, the correlation is said to be "multiple." For example, if we want to study the relationship between the amount of fertilizer and rain used and how much wheat is grown per acre, this is an example of a problem with multiple correlations.

In a partial correlation, on the other hand, we look at more than two variables, but we only look at the two that affect each other in a way that doesn't change the effect of the other variable. For example, in the previous example, if we look at the relationship between yield and fertilizers used during times when a certain average temperature was present, this is an example of a partial correlation problem.

1. Linear AndNonlinear (Curvilinear) Correlation:

Whether there is a linear or non-linear relationship between the variables depends on whether the ratio of change between the variables stays the same. A correlation is said to be linear when the ratio of how much one variable changes to how much another variable change stays the same.

For example, based on the values of the two variables below, it is clear that their rate of change is the same:

X: 10 20 30 40 50

Y: 20 40 60 80 100

Non-linear or curvilinear correlation is when the amount of change accordingly.

For example, doubling the amount of fertilizers would not necessarily double the amount of wheat that could be grown. So, these are the three most important types of correlation based on how the variables move, how many of them there are, and how much they change.

2. Positive And Negative Correlation

The direction of change determines whether the relationship between the variables is positive or negative. A positive correlation develops when two variables consistently move in the same direction. A negative correlation is observed when both variables demonstrate reversed movement. This indicates that while one variable increases, the other variable drops, and vice versa.

5.5 Summary

Correlations are a good way to describe simple connections between data. Say, for exampleyou are examining a dataset regarding campgrounds in a mountainous park. You are inquiring about the potential correlation between the elevation of the campsite, which refers to its height above sea level, and the average high temperature throughout the summer season.

- There are two ways to measure each campsite: altitude and temperature. When you use a correlation to compare these two things across your sample, you can find a linear relationship: as elevation goes up, temperature goes down. They go against each other.
- The presence of a linear or non-linear relationship between the variables is determined. A correlation is considered linear when the ratio of the change in one variable to the change in another variable remains constant.
- The user's text is a bullet point symbol. The direction of change determines whether the link between the variables is positive or negative. A positive correlation develops when two variables consistently move in the same direction. A negative correlation is

observed when both variables demonstrate an opposite or inverse relationship. This suggests a negative correlation between the two variables, meaning that as one variable increases, the other variable drops, and vice versa.

5.6 Self-Assessment Questions

1. What does correlation mean?
2. What is correlation and types?
3. Why is correlation important?
4. What are the 4 types of correlation?
5. What are the benefits of correlation in statistics?
6. What are the advantages and disadvantages of correlation?
7. What is correlation in statistics?
8. What is the difference between cause and effect and correlation?
9. How to calculate linear correlation?
10. What is linear and non-linear correlation?

5.7 References / Reference Reading

- McCarney R, Warner J, Iliffe S, van Haselen R, Griffin M, Fisher P (2007). "The Hawthorne Effect: a randomized, controlled trial". BMC Med Res Methodol. 7 (1): 30.
- Rothman, Kenneth J; Greenland, Sander; Lash, Timothy, eds. (2008). "7". Modern Epidemiology (3rd ed.). Lippincott Williams & Wilkins. p. 100.
- Mosteller, F.; Tukey, J.W (1977). Data analysis and regression. Boston: Addison-Wesley.

CHAPTER 6 — Causation and Correlation Analysis

Leaning Objective:

After studying the unit, students will be able to:

- Know what causation is.
- Understand correlation and causation analysis
- Explain scatter diagram and Pearson's coefficient analysis

Structure

6.1 Correlation and Causation

In statistics, two or more variables are said to be related if their values change in the same way: if the value of one variable goes up or down, so does the value of the other variable (although it may be in the opposite direction).

For example, there is a relationship between "hours worked" and "income earned" if the number of hours worked goes up and the amount of money earned also goes up. If we look at the two variables "price" and "buying power," we can see that as the price of goods goes up, a person's ability to buy these goods goes down (assuming a constant income).

But just because there is a link between two variables doesn't mean that the change in one variable is the cause of the change in the other variable.

The word "causation" means that one event happened because of the other event, or that there is a link between the two events. This idea is also called "cause and effect."

In theory, distinguishing between the two sorts of partnerships is a straightforward task. An action or event can either lead to another (such as smoking increasing the likelihood of developing lung cancer) or just be associated with another (for instance, smoking is correlated with alcoholism, but it is not a direct cause of alcoholism). However, in reality, demonstrating causation is still challenging compared to demonstrating correlation.

Objectives

The primary objective of many studies and scientific analysis is to determine the correlation between two variables. For example,

- Is there a correlation between an individual’s level of education and their overall health?
- Does owning a pet correlate with increased longevity?
- Did the implementation of a company's marketing effort result in a significant boost in their product sales?

With these and other questions, we are trying to find out if there is a relationship between the two variables. If there is a relationship, this may help us figure out if one thing causes the other.

Way to Measure

A statistical correlation between two variables is quantified using a Correlation Coefficient, which is a singular numerical value that indicates the degree of relationship between the two

variables. The symbol (r) represents this number.The coefficient is a numerical value ranging from +1 to -1, indicating the strength and direction of the association.
If the correlation coefficient has a value below 0, it means that the two variables are linked in a bad way.

Limitations:

Correlation coefficients are typically used to quantify the strength of a linear association. For instance, a skilled worker who invoices based on an hourly rate exhibits a linear correlation between the number of hours worked and the corresponding earnings. This is due to the fact that their revenue increases uniformly with each additional hour worked. Alternatively, if the tradesperson imposes an initial call-out fee and a decreasing hourly rate as the task progresses, the association between the number of hours done and income would not follow a linear pattern, resulting in a correlation coefficient that is likely closer to 0.

It is important to exercise caution while determining the significance of the variable "r". It is feasible to establish connections across many entities; nevertheless, these connections may be attributed to factors unrelated to the two entities under examination.
For instance, the sales of ice cream and sunscreen exhibit a cyclical pattern throughout the year, with regular fluctuations. However, this correlation is attributed to the seasonal impacts (e.g., increased sunscreen usage and ice cream consumption during hot weather) rather than a direct causal relationship between sunscreen sales and ice cream sales.

The correlation coefficient should not be employed to make causal inferences. By examining the magnitude of the correlation coefficient "r," we can infer a relationship between two variables. However, the value of "r" does not provide information on the direction or causality of the relationship.

6.2 Way to Establish

Causality is an area of statistics that is often misunderstood and abused by people who think that just because there is a correlation between two things, there must be a cause-and-effect relationship.

The best way to show that one variable leads to another is through a controlled study. In a controlled study, the population or sample is split into two groups that are almost identical in every way. Then, each group gets a different treatment, and the results are measured for each.

There are ethical limits to how controlled studies can be used. For example, it wouldn't be right to use two similar groups and make one of them do something harmful while the other doesn't. To get around this, correlation and causation for the population of interest are often studied through observational studies. The studies can look at how the groups act and what happens, as well as any changes that happen over time.

The goal of these studies is to provide statistical information that can be added to other kinds of data that would be needed to figure out whether or not there is a cause-and-effect relationship between two variables.

6.3 Scatter Diagram

A scatter diagram visually represents the values of two variables, X and Y, and demonstrates the correlation between these two variables. The X variable values are represented on the horizontal axis, while the Y variable values are represented on the vertical axis. The assignment of variables as X or Y is irrelevant when constructing a scatter plot and computing the correlation coefficient.

In the regression model, one variable is designated as the independent variable, while the other is designated as the dependent variable. The correlation methods are same for both variables, and it is not possible to establish causation or discern the direction of influence.

When to use a scatter diagram

When you have a set of numerical data that is grouped together in pairs.

When your variable that is being measured can have many values for each value of the variable that is being manipulated.

When attempting to determine the relationship between two entities, such as:

When attempting to ascertain the potential cause of an issue,

When determining if two correlated effects are attributable to a common cause.

Prior to creating a control chart, it is crucial to examine for autocorrelation.

While a scatter diagram may indicate a correlation, it is important not to infer causation between the variables. Both entities could be influenced by a third factor.

When the data are put on a graph, the relationship is stronger if the diagram looks like a straight line. If a line isn't clear, statistics (N and Q) help figure out how likely it is that there is a relationship. If the numbers show that there is no link, the pattern could have happened by chance. If the scatter diagram doesn't show any connection between the variables, you might want to think about whether the data could be separated into groups.If the diagram doesn't show a relationship, think about how much the independent (x-axis) variable changed. Sometimes it's hard to see a connection because the data don't cover a wide enough range.

6.4 Pearson's Coefficient of Correlation

It is a test statistic that shows how closely two continuous variables are related statistically. Because it is based on the method of covariance, it is known as the best way to measure how two important variables are linked. It tells how strong the relationship is and which way it goes.

Assumptions

- Case independence: Cases should be mutually unaffected.
- Linear relationship: The correlation between two variables should be a straight line.
- Homoscedasticity refers to the property where the scatterplot of residuals has a rectangular shape.

Properties

- Pure number: It doesn't matter what unit is used to measure it. For example, if the unit of measurement for one variable is inches and the unit of measurement for the other variable is quintals, the value of Pearson's correlation coefficient does not change.

- Correlation between two variables is symmetric if the coefficients of the two variables are the same. This means that the coefficient value will stay the same between X and Y or Y and X.

Degree of Correlation

- Perfect: A perfect correlation occurs when the value is close to 1. In this case, when one variable increases, the other variable tends to increase (if positive) or decrease (if negative).
- High degree: A connection is considered strong when the coefficient value is within the range of 0.50 to 1.

- A moderate degree: A number ranging from 0.30 to 0.49 is referred to as a medium correlation.
- Low degree: A correlation is considered modest if its value is below +0.29.
- No correlation: When the value is exactly zero.

Calculation of Pearson Correlation Coefficient

A Pearson Correlation Coefficient quantifies the linear relationship between two variables. The value always falls within the range of -1 and 1, inclusive.

- "A value of -1 indicates the absence of a linear relationship between the two variables."
- "A value of 0 indicates the absence of a linear relationship between the two variables."
- "A value of 1 indicates a perfect linear relationship between the two variables."

A Pearson Correlation Coefficient, written as r, can be found by using the following formula:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Steps-

Step 1: Calculate the Mean of X and Y

Step 2: Calculate the Difference between Means

Step 3: Calculate the Remaining Values

Step 4: Calculate the Sums

Step 5: Calculate the Pearson Correlation Coefficient

6.5 Summary

1. Correlation analysis quantifies the degree to which one variable is affected by variations in another one. A strong correlation between two variables indicates a robust link. A low correlation signifies a weak connection between two variables.
2. Market analysts employ correlation analysis to scrutinize the quantitative data acquired from sources like surveys and live polls. They examine patterns, significant correlations, and trends between two variables or sets of data.

3. It is a statistical measure that quantifies the extent of linear relationship between two variables, indicating the degree to which they move together at a consistent rate. A common approach of discussing direct relationships without openly addressing causality.

6.6 Self-Assessment Questions

4. What is a scatter diagram?
5. Explain analysis in scatter diagrams with examples.
6. Differentiate between scatter diagram and pie chart.
7. What are the 4 types of correlation?
8. What is Pearson’s coefficient correlation?
9. What are the advantages and disadvantages of Pearson’s model of correlation?
10. What is correlation in statistics?
11. What is the difference between cause and effect and correlation?
12. How do you draw a Pearson correlation graph?
13. What does the Pearson's correlation coefficient show?

6.7 References / Reference Reading

- McCarney R, Warner J, Iliffe S, van Haselen R, Griffin M, Fisher P (2007). "The Hawthorne Effect: a randomized, controlled trial". BMC Med Res Methodol. 7 (1): 30.
- Rothman, Kenneth J; Greenland, Sander; Lash, Timothy, eds. (2008). "7". Modern Epidemiology (3rd ed.). Lippincott Williams & Wilkins. p. 100.
- Mosteller, F.; Tukey, J.W (1977). Data analysis and regression. Boston: Addison-Wesley.

CHAPTER 7 — Regression Analysis

Learning Objective:

After studying the unit, students will be able to:

- Know what regression analysis
- Understand the meaning of regression coefficient
- Explain the relation between regression and causation

Structure

7.1 Regression Analysis

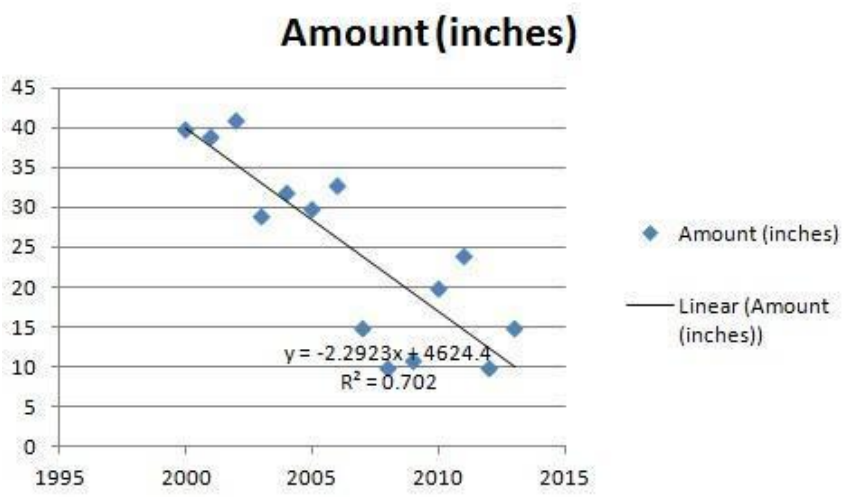
It is a statistical method to measure and describe the strength and nature of the linkage.

Linear regression is the most prevalent variant of this method. It can also be called simple regression or ordinary least squares (OLS). Linear regression utilises a line of optimal fit to ascertain the linear association between two variables. Linear regression on a graph is depicted as a straight line, with the slope indicating the magnitude of the effect of a change in one variable on the other.

It is a useful technique for discovering associations between variables in data, but it is not suitable for establishing causation. It is utilised in many roles within the domains of business, finance, and economics. For example, it is used to help investment managers determine the value of assets and analyse the effects of factors such as commodity prices.

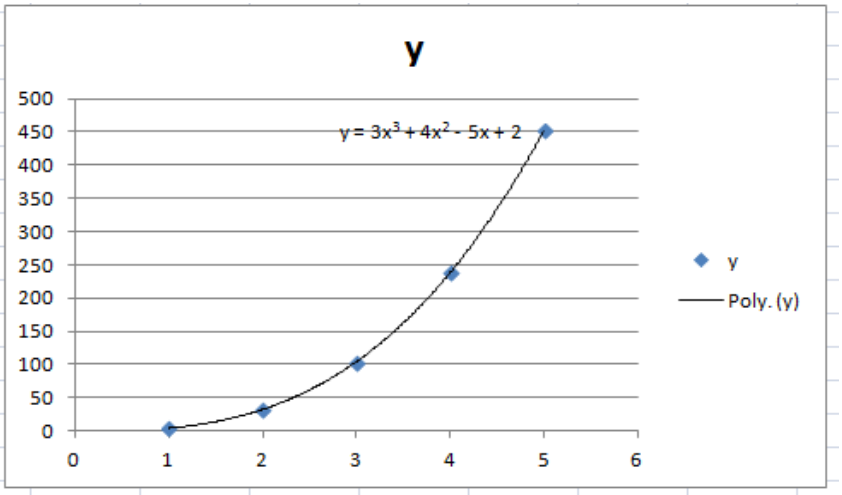
7.2 Regression Equations and Estimation

In the field of statistics, a regression equation is employed to determine the presence of a relationship between two sets of data. For example, if you do an annual measurement of a child's height, you may observe an average annual growth of approximately 3 inches. A regression equation can be employed to demonstrate the veracity of this pattern, which entails an annual growth of three inches. Indeed, nearly all phenomena in the tangible realm, ranging from the cost of fuel to the occurrence of storms, may be accurately represented using mathematical equations. This enables us to forecast future events.



An equation is 5employed to demonstrate the regression line. The equation for this particular example can be expressed as -2.2923x + 4624.4. Graphing the equation -2.2923x + 4624.4 would result in a line that closely approximates your data.

It is rare for all of the data points to align perfectly on the regression line. The dots in the image displayed above exhibit a slight dispersion in relation to the line. All the dots in the following image are positioned on the line. Polynomial regression utilises a polynomial equation to accurately represent the relationship between the points, resulting in a tilted line.



7.3. Regression Coefficient

- It is usually written as "b."
- It is written out as an original unit of data.
- If there are two variables, x and y, there will be two separate values for the regression coefficient. One answer can be found by treating x as the independent variable and y as the dependent variable, while the other solution can be found by treating y as the independent variable and x as the dependent variable. The coefficient byx represents the relationship between y and x, while the coefficient bxy represents the relationship between x and y.
- If one regression coefficient exceeds unity, then the remaining coefficients will be below unity.
- The correlation coefficient is equal to the geometric mean of the two regression coefficients. The value of R is equal to the square root of the product of byx and bxy.
- The expression (byx + bxy)/2 is bigger than or equal to r.

- The regression coefficients stay the same even if the starting point changes. But they don't have anything to do with how the scale changes. It means that taking any constant away from the value of x and y won't change the regression coefficients. Any constant multiplied by x and y will change the regression coefficient.

7.4 Relationship Between Correlation and Regression Coefficients

Correlation and regression are used to describe some kind of relationship between two numbers that are thought to be linked in a straight line. In this article, well learn more about these topics, find out what the difference is between correlation and regression, and look at some examples related to these topics.

Correlation and regression are both statistical measurements used to show how two variables are related. For example, suppose a person is driving an expensive car then it is assumed that she must be financially well. Correlation and regression are used to put a number on this relationship.

Correlation Definition

It is a numerical measure for evaluating the association between variables. If a modification in one variable results in a proportional modification in another variable, the two variables are deemed to be directly linked. Likewise, two variables are said to have an inverse relationship if a change in one variable leads to a commensurate change in the other variable, or vice versa. If manipulating an independent variable does not lead to any change in the dependent variable, it can be inferred that there is no correlation between the two variables. Correlation can appear as positive (direct correlation), negative (indirect correlation), or zero. The correlation coefficient measures the strength of this relationship.

Regression Definition

It is a statistical method used to quantify the impact of a change in one variable on another one. It is employed to determine the correlation between two variables and identify the causal relationship between them.

Standard Error of Estimate

It represents the mean deviation of the observed values from the regression line. The units of the response variable provide a measure of the average error of the regression model. When

the value is small, it indicates that the observations are in close proximity to the fitted line.

Instead of relying on R squared, the standard error of the regression can be utilised as a measure of the accuracy of the predictions. When using a regression model for predictions, it is more crucial to consider the standard error of the regression rather than focusing on R-squared.

7.5 Summary

- In the field of statistics, the relationship between two or more variables is established when their values exhibit a similar pattern of change.
- A statistical correlation between two variables is quantified by a Correlation Coefficient, which is a singular numerical value that indicates the degree of relationship between the two variables. The symbol (r) represents this number.
- The coefficient is a numerical value ranging from +1 to -1, indicating the strength and direction of the association.
- A correlation coefficient below 0 indicates a negative relationship between the two variables. This implies that the variables exhibit inverse relationships.
- Causality, a branch of statistics, is frequently misconstrued and misused by individuals who mistakenly assume that a correlation between two variables implies a cause-and-effect connection.
- Controlled research is the most effective method for demonstrating a causal relationship between two variables. In a controlled study, the population or sample is divided into two groups that are nearly identical in all aspects. Subsequently, each group is subjected to a distinct intervention, and the outcomes are quantitatively assessed for each group.
- It is a statistical measure that quantifies the degree of association between two continuous variables. Due to its reliance on the covariance approach, it is widely regarded as the most effective means of quantifying the relationship between two significant variables. It assesses the strength of the relationship and its direction.

7.6. Self-Assessment Questions

1. How do you find the correlation of a scatter diagram?
2. How do you calculate the Pearson's correlation coefficient?
3. What type of correlation does the scatter plot show?

4. How is a scatter plot used to represent correlations between two variables?
5. What is a regression analysis in statistics?
6. What is regression analysis and its types?
7. Why is regression analysis used?
8. How do you calculate regression?
9. How do you find P value in regression?
10. What is the Relationship between Correlation and Regression coefficients?

7.7. References

- McCarney R, Warner J, Iliffe S, van Haselen R, Griffin M, Fisher P (2007). "The Hawthorne Effect: a randomized, controlled trial". BMC Med Res Methodol. 7 (1): 30.
- Rothman, Kenneth J; Greenland, Sander; Lash, Timothy, eds. (2008). "7". Modern Epidemiology (3rd ed.). Lippincott Williams & Wilkins. p. 100.
- Mosteller, F.; Tukey, J.W (1977). Data analysis and regression. Boston: Addison-Wesley.

CASE STUDY

Find the two regression equations of X on Y and Y on X from the data given below, taking deviations from the actual means of X and Y.

Price(Rs.)	10	12	13	12	16	15
Amount demanded	40	38	43	45	37	43

Unit IV: Index Numbers

CHAPTER 8 — Introduction to Index Numbers

Learning Objective:

- After studying the unit, students will be able to:
- Know what index numbers are.
- Understand the uses of index numbers Explain the advantages and limitations of index numbers

Structure

8.1. Meaning and Uses of Index Numbers

Index numbers are the most important thing you need to know about statistics. Consider the implications of not having these numerical values when altering a variable in a statistical analysis. The procedure itself will ultimately prove ineffective. Index numbers are a quantitative method used to assess changes in one or more variables within a specific timeframe. These statistics indicate a shift in the overall state of affairs, rather than a quantifiable value. The % form is utilized to express an index number.

8.2. Importance of Index Number

Index numbers are predominantly utilized for analyzing the economic conditions of specific regions.

So, Index numbers are important because they can be used to measure how much the economy has changed over a certain time period. It helps us figure out how these changes affect things we can't measure directly.

8.3. Features and Characteristics of Index Numbers

The most important things about index numbers are as follows:

- A weighted average is employed to assess relative changes in situations where it is not possible to obtain precise data.
- The index number provides an indication of potential variations in aspects that may not be readily quantifiable. It provides a general understanding of the extent to which things have changed.
- The measurement of an index number varies depending on the linked variable being considered.
- It facilitates the comparison of the levels of a phenomenon on a specific day with those on a prior date.
- It serves as an illustrative instance of a specific type of average known as a weighted average. Index numbers are universally applicable. The identical index employed to ascertain price fluctuations can also serve as a metric for gauging.

8.4. Classification of Index Numbers

There are different kinds of index numbers that are used for different things. We will look at the different kinds of Index numbers to learn the same thing. This part about the different

kinds of Index numbers will help students understand how important each kind is for the task they are practicing for.

Value Index

It is made by dividing the sum of all the values for a certain period by the sum of all the values for the base period. The value index is used, among other things, for inventory, sales, and international trade.

Quantity Index

It is used to track changes in how much of a good is made, used, and sold during a certain time period. It shows how the amounts of goods have changed over a certain time period. “One example of a Quantity Index is the Index of Industrial Production (IIP).”

Price Index

A price index number is a way to track how prices change over time. It will show the relative value, not the value in and of itself. “Price indices include the Consumer Price Index (CPI) and the Wholesale Price Index (WPI)”.

Uses

We knew what the Index numbers were made of and what kinds they came in. W we now talk about how Index numbers are used.Index numbers are useful in many studies, from the simplest to the most complicated. For example, it is used to get a general idea of how many people live in a country and how fast rare animals are dying out in a certain area. Index numbers can be used in many more ways, let's find out:

- It helps measure changes in the cost of living and the standard of living.
- Price level changes are taken into account when setting wage rates. When price levels are known, wage rates could be changed.
- The index number of prices is used to decide how the government will run. Index numbers are what make it possible for fiscal and economic policies to keep prices stable.
- It gives a guideline for comparing different economic factors between countries, such as the standard of living in each country.

Advantages of Index Number

Index numbers have benefits that are directly linked to how they are used. So, to sum up the benefits, they are as follows:

- It makes changes to primary data at different costs, which helps deflate.

- Index numbers are used a lot in economics and help come up with the right policies. These kinds of results also help researchers start their work.
- It helps when there is a trend, like when you want to figure out what will happen with irregular forces and cyclical forces.
- Index numbers can be used to predict how things in the economy might change in the future. This time series analysis is used to find trends and changes that happen in cycles.
- The number can be used to track changes in the standard of living in different countries over a certain amount of time.

Diadvantages

We know that everything has both good points and bad points. Index numbers have a lot of good points, but this is where some of their flaws start to show. These are the things that can't be done with index numbers:

- Since index numbers are based on samples, there is a chance that they will be wrong. These samples are put together after a lot of thought, so mistakes are possible. You can also find it in weights or base periods, among other places.
- It is always worked out based on the things. Items that are chosen in this way might not really be trendy, which makes the analysis wrong.
- Index numbers can be made in many different ways. Because there are so many ways to do things, the results may show different sets of values, which can add to the confusion.
- The index numbers give a general idea of how the changes affect each other. Changes in variables that are looked at over a long period of time may also not be reliable.
- The way the representative goods are chosen could be off. Since these goods are based on samples, it makes sense.

8.5 Summary

- Index numbers are a key notion in statistics that one must grasp. Reflect about the consequences of not possessing these numerical values when modifying a variable in a statistical analysis. The technique will ultimately be useless. Index numbers are a quantitative technique employed to evaluate alterations in one or more variables

throughout a particular period of time. These data suggest a change in the general situation, rather than a measurable assessment. The percentage form is used to denote an index number.

- The user's text consists of a bullet point symbol.Index statistics are mostly used to analyze the economic situations of particular regions. The index number denotes the relative magnitude.
- The value index number is computed by dividing the aggregate sum of values for a given time frame by the aggregate sum of values for the reference period. The value index is employed for multiple objectives.
- It is used to track changes in the production, consumption, and sales of a certain commodity over a defined period of time. It demonstrates the variations in the quantity of goods throughout a certain period of time.
- Index numbers are highly important in a broad spectrum of studies, ranging from simple to complex. For example, it is used to gain a comprehensive picture of the population size in a country and the rate at which endangered animals are decreasing in a certain region.

8.6 Self-Assessment Questions

1. Give the uses of index?
2. Define the meaning of number index?
3. Mention the uses of index numbers in statistics?
4. What can be the uses and limitations of index number?
5. Discuss the advantages of index numbers?
6. Discuss the steps in the construction of an index number?
7. Discuss the uses of construction of index numbers?
8. Discuss the two methods of constructing an index number?
9. What is the formula of index number?
10. What are the different types of index numbers?

8.7 References / Reference Reading

- Cohen, Jerome B. (December 1938). "Misuse of Statistics". Journal of the American Statistical Association. JSTOR. 33 (204): 657–674.
- Freund, J.E. (1988). "Modern Elementary Statistics". Credo Reference.

- Huff, Darrell; Irving Geis (1954). How to Lie with Statistics. New York: Norton. The dependability of a sample can be destroyed by [bias]... allow yourself some degree of skepticism.
- Nelder, John A. (1999). "From Statistics to Statistical Science". Journal of the Royal Statistical Society. Series D (The Statistician).

CHAPTER 9 — Construction of Index Numbers

Learning Objective:

After studying the unit, students will be able to:

- Know what construction of index number is
- Understand the methods for construction
- Explain the difference in each methodology

Structure

9.1. Types of Methods of Construction

Construction of Index Numbers

Index numbers can be generated by many methodologies. Usually, there are two approaches to producing an index number: the "simple" approach and the "weighted" approach. In addition, the easy technique can be classified into two forms-

Simple Approach - Aggregative and Relative

Weighted Method - Aggregative and Relative

Now, let's examine the functioning of each of these methods for constructing an index.

9.2. Simple Aggregative Method

We use this way of building to figure out the index price. So, the ratio of the total cost of a good in a given year to its total cost in the base year is shown as a percentage.

Simple Aggregative Price Index ❑ (ΣPn/Σ PO) * 100 Where.

ΣPn = Sum of the entire respective commodity.

ΣPo = Sum of the entire respective commodity in base period.

It's easy to understand how the simple aggregative index works. But this method has a big flaw that makes it very bad.

Moreover, altering the units in any manner will result in a corresponding modification of the index number. When absolute amounts are used, the situation is completely reversed. Therefore, it is advisable to consider distinct values for each of the three years.

Simple Average of Relatives

A replacement would be better than a simple average index if you want to get rid of the mistakes and flaws that come with it. So, to build the Index, we can use a simple method based on the average of relatives.

By employing this technique, we may convert the actual value of any variable into a proportion relative to the base period, expressed as a percentage. The term used to refer to these quantities is "relatives. Therefore, the index numbers that we obtain are likely to remain unchanged.

9.3. Weighted Method

It is crucial to address the requirements of any basic or non-weighted approaches. Therefore, in this scenario, we employ any criterion that we deem appropriate to determine the worth of any commodity. Typically, this factor is determined by the selling price in the initial year. The indexes are categorized into the following groups:

- Weighted Aggregative Index
- Weighted average of relatives.

Let us closely examine the following two indices.

Weighted Aggregative Index Method

Most of the time, this is how we figure out the price of a good. A very rough factor is used to figure out how much something weighs. Most likely, these things will be different and can be anything. It can be a number or the number of units that are sold during the base year.

The year doesn't have to be the base year. It could be an average of other years or just any year. Well, the choice of it will depend entirely on how important the year is. So, besides the amount, it's up to us to decide how important a certain year is.

Most of the time, the Weighted Aggregative Index is given as a percentage. Because of this, there are different ways to do the same thing. Here are some of them:

1. **Laspenyrs, Index**

Determined as

Formula = (ΣPnQo/ΣpoQn)*100

2. **Paasche's Index**

This refers

Formula = (Σ PnQn/ΣpoQn)*100

3. **Some of the methods that depend on a typical time period:**

Index (ΣPnQt/ΣpoQn)*100, here, the subscript "t" symbolizes the typical period of time in years.

The quantities of these years are the values of weight

Marshall-Edgeworth Index

Inthis we nclude booth current and base year

MarshaIL Edgeworth Index -

$$[\sum Pn(Q\,Qn)/\sum Po(Q\,Qn)] * 100$$

4. Fisher's Ideal Price Index

The geometric mean of Laspeyres' and Paasche's is the Fisher's Ideal Price Index.

$$\text{Formula} - \sqrt{[(\sum PnQo/\sum PoQo)*(\sum PnQn/\sum PoQn)]} * 100$$

Average of Relatives by Weight

We avoid the problem that comes with the simple average method by using the weighted average of relatives. Also, the preferred method is the weighted geometric mean, but sometimes the weighted arithmetic mean is used. So, this is how the weighted AM looks when the values of the base year weights are used:

Formula = (Σ PnQo/ΣPoQo)*100

Base Shifting

It is often necessary to change the reference base of an index number series from one time to the next without going back to the original raw data and recalculating the whole series. Most of the time, this change in the base period is called "shifting the base." There are two major reasons why the base should be moved:

1. The old base is too old and can't be used to compare things as well as it used to. By changing the base, the series can be written about a more recent time period.
2. It may be useful to compare several index number series that have been compared on different base periods. This is especially true if the different series are to be shown on the same graph. This could mean that the base period needs to change.

When changing the base period, one option is to recalculate all index numbers with the new base period. A simpler way to get a close estimate is to divide all of the index numbers for the different years in the old base period by the index number for the new base period and write the result as a percentage. These numbers are the new index numbers, with 100% being the index number for the new base period.

Mathematically, this method can only be used exactly if the index numbers pass the circular test. But, luckily, for many types of index numbers, the method gives results that are close enough to what would be expected theoretically.

Base Shifting and Splicing are two parts of index numbers that are used to change the reference base of an index number and join two different index numbers together.

Chain Base Shifting

A chain base index number is one where the numbers for each year are first shown as a percentage of the year before. These people are called "Link Relatives." Then, we need to link them together by multiplying them one by one to make a chain index. In this method, the base year changes every year, which is different from methods with a fixed base. So, it will be 2000 for the year 2001, 2001 for the year 2002, and so on.

Advantages of Chain Index Numbers Method

1. This method makes it possible to add new items to the series and get rid of ones that are no longer needed.
2. It doesn't change with the seasons.
3. It allows the weights to be changed as often as possible.
4. It is a very useful way to compare economic and business data from one time period to another.
5. In a business, management usually compares the current period to the one right before it, not to any other time in the past. The base year changes every year with this method, which makes it more useful for management.

Disadvantages of Chain Index Numbers Method

1. It is hard and takes a lot of time.
2. The numbers in the link are a percentage of what they were the year before. Chained percentages are not a good way to compare things over a long period of time.
3. With this method, if we don't have the data for any one year, we can't figure out the chain index number for the next year. This is because we need to figure out who the link's relatives are, which we can't do in this case.
4. If there is a mistake in the calculation of any of the link relatives, the mistake is added to and all of the other link relatives are also wrong. So, the whole series will give a wrong idea of what is going on.

Difference between Fixed Base Index and Chain Base Index

In a fixed base index, the value of one year is used as the base to make indices for different years. The base year stays the same for all of the index series.

In the chain base index, each year's numbers are given as a percentage of the year before. Then, these percentages are linked together by multiplying them one by one to make a set of chain indices. The base year for chain base index changes from year to year.

Conversion of Chain base Index number to Fixed base Index number
Current year FBI = (Current year CBI x Previous year FBI)1100.
The fixed base index for the first year will be taken the same as the chain base index

Splicing

When the base of an index number series gets too old, it is sometimes stopped from being built. Some of the most recent years could be used to make a new set of index numbers. You might want to set the new indices to match the old ones. Splicing is the statistical process of putting together an old index number series.

Occasionally, it is logical to shift the base period of an index number series to a more current time. Over time, certain commodities used to construct indices may be substituted with new commodities, leading to changes in their respective weighting. Occasionally, the weights may become inaccurate, necessitating the utilisation of the updated weights. Due to unknown circumstances, the index number series has seen a disruption in continuity. As a result, we now have two distinct index number series with different base periods that cannot be directly compared. Therefore, it is crucial to merge these two distinct sets of indices into one cohesive and uninterrupted set. "Splicing" is the statistical term used to describe the process of joining two series of indices in order to create a seamless flow between them. So, splicing means turning two sets of index numbers with different base periods that overlap into a single set of index numbers. In the form of an equation, we can say,

$$\text{Spliced Index Numbers} = \text{New Index No. of current period} \times \frac{\text{Old index No. of new base period}}{100}$$

The steps for splicing are shown in the following example:

Splicing the New Series of Indices with the Old Series of Indices

Year	Consumer Price Index (1990 = base) (Old Index No. series)	Consumer Price Index (1994 =base) (New Index No. series)	Spliced Consumer Index [New index (114/100)]
1990	100		100
1991	110		110
1992	108		108
1993	114	100	100 (114/100) = 114
1994		108	108 (114/100) = 123
1995		116	116 (114/100) = 132
1996		112	112 (114/100) = 128

In the example given above, the old series ended in 1993 and the new series began in the same year. As shown in Column No. 4, the new series began with 1993 as its base year.

Conversely, the previous index number series might be combined with a fresh index number series. This implies that rather than progressing with older series, new ones could be regressed. In order to accomplish this task, I will provide you with the formula.

$$\text{Old Index No. of current period} \times \frac{100}{\text{Old index No. of new base period}}$$

Under this approach (Splicing the old series with the new series) the spliced

indices are as follows:

Year	1990	1991	1992	193	1994	1995	1996	1997	1998
Spliced Index	88	96	97	95	100	108	116	112	120

Deflating

As we all know, the price of goods and services goes up over time. As a result, the value of money (how much it can buy) goes down. Because of this, the real wage falls below the money wage. In this case, the real wage can be found by lowering the money wage by the amount that prices have gone up. So, "deflating" is the process of figuring out the real wage by putting the money wages through the right price indices to account for changes in the price level. The following formulas can be used to describe this process:

$$\text{Real wage} = \frac{\text{Money wage}}{\text{Price index}} \times 100, \text{ and}$$

$$\text{Real wage index number} = \frac{\text{Current period's real wage}}{\text{base period's real wage}} \times 100$$

Let's use the following data about wages and the price index from different years as an example. It would illustrate the procedure of constructing real wage index numbers.

Construction of Real Wage Index

Year	Wages (Rs.)	Price index	Real wage (Rs.) (deflated income)	Real wage index (1995=100)
1995	200	100	(200/100) 100=200	(200/200) 100 = 100
1996	280	130	(280/130) 100 = 215	(215/200) 100 = 107.5
1997	280	190	(280/190) 100 = 147	(147/200) 100 = 73.5
1998	360	240	(360/240) 100 = 150	(150/200) 100 = 75
1999	390	280	(390/280) 100 = 139	(139/200) 100 = 69.5
2000	420	280	(420/280) 100 = 150	(150/200) 100 = 75

9.4 Difficulties in the Construction of Index Numbers

There are a lot of problems with making index numbers. Here are some of the problems that come up when making index numbers:

1. **Difficulties in Choosing a Base Period:**

The first problem is figuring out what year to use as a starting point. The first year must be the same. But it's hard to figure out a causal year. Also, after a certain amount of time, a normal year is no longer normal. So, it's not a good idea to use the same base period for several years.

2. Problem in Commodity Selection:

Another problem is choosing the representative commodities for the index number. The choice is not an easy task. They have to choose from many different things that people eat. Changes in how people buy things could make the number of indexes useless. Because of this, it is not easy to choose goods that are representative.

3. Problems in Price Compendium:

Ensuring adequate and appropriate nutrition is another challenge. Obtaining anything consistently from a single source is not always possible. Furthermore, the issue of establishing intermediate pricing arises. There is a wide range of retail prices. Index numbers are utilised to determine wholesale prices.

4. Difficulty in Choosing a Statistical Approach:

Another problem is figuring out the best way to calculate averages. But each strategy leads to a different set of results. Because of this, it's hard to decide which strategy to use.

5. Difficulties Resulting from Changes Over Time:

Because technology keeps getting better, goods are always changing in the world we live in now. So, people start buying them, and the old goods are replaced with new ones. Also, changes in commodity prices may be caused by changes in technology. They could fall down. But when the index numbers are calculated, new goods are not added to the list of goods. So, the index numbers that are based on old goods are not real.

6. It is not Possible to Make a Comparison:

Index numbers don't let you compare prices from different countries. The goods consumed and used to figure out an index number vary from country to country. In advanced countries, price indices include things like meat, eggs, cars, and electronics, but they don't in backward countries. In the same way, goods have different weights. This makes it impossible to compare index numbers from different countries.

7. It is not possible to make comparisons between different locations:

Even if different places in the same country are chosen, they can't all have the same index number. This is because people have different eating habits. People in the north of India buy different things than people in the south. So, it would be wrong to give both the same index number.

8. Not appropriate to individuals:

An index number can't be used by just one member of the group for which it was made. If the price level index number goes up, it might not affect a person. This is because an index number shows how things are on average.

Precautions to Be Taken While Constructing an Index Number

The first step is to get an estimate of how much the chosen goods will cost. We all know that the prices of different goods vary from one place to another and even between stores in the same market. It's also important to decide whether the prices should be wholesale or retail. The choice would depend on what the index number would be used for.

The choice of a good base year is another safety measure taken when making an index number. The base year is used to compare different years. As the base year, a normal year should be used. It shouldn't have any weird things happen, like wars, earthquakes, or other natural disasters.

The next thing to think about when building an index number is what it will be used for. When a value index is needed, it is right to figure out a volume index. With the help of index numbers, the goods are chosen and their prices are set.

These are the steps to take to deal with the problems that come up when making index numbers.

9.5. Summary

- Index numbers can be generated by many techniques. Generally, there are two approaches to producing an index number: the "simple" approach and the "weighted" approach. In addition, the basic technique can be classified into two specific types:

simple aggregative and simple relative. There are two types of weighted methods: weighted aggregative and weighted average or relative.

- A fixed base index uses the value of one year as the reference point to create indexes for subsequent years. The base year remains constant for all index series.
- In the chain base index, each year's numbers are given as a percentage of the year before. Then, these percentages are linked together by multiplying them one by one to make a set of chain indices. The base year for chain base index changes from year to year.
- When the base of an index number series gets too old, it is sometimes stopped from being built. Some of the most recent years could be used to make a new set of index numbers. You might want to set the new indices to match the old ones. Splicing is the statistical process of putting together an old index number series.
- As we all know, the price of goods and services goes up over time. As a result, the value of money (how much it can buy) goes down. Because of this, the real wage falls below the money wage. In this case, the real wage can be found by lowering the money wage by the amount that prices have gone up. So, "deflating" is the process of figuring out the real wage by putting the money wages through the right price indices to account for changes in the price level.

9.6. Self-Assessment Questions

1. What are the difficulties in constructing the index numbers?
2. Discuss the nature of an index number?
3. What is splicing and base shifting?
4. What do you mean by base conversion and base shifting?
5. What are the types of splicing in statistics?
6. What do you mean by fixed base index number?
7. What is chain base?
8. What are the advantages of chain base index?
9. How do you calculate chain index?
10. What is Fisher's ideal index?

9.7. Reference Readings

- Cohen, Jerome B. (December 1938). "Misuse of Statistics". Journal of the American Statistical Association. JSTOR. 33 (204): 657–674.
- Freund, J.E. (1988). "Modern Elementary Statistics". Credo Reference.
- Huff, Darrell; Irving Geis (1954). How to Lie with Statistics. New York: Norton. The dependability of a sample can be destroyed by [bias]... allow yourself some degree of skepticism.
- Nelder, John A. (1999). "From Statistics to Statistical Science". Journal of the Royal Statistical Society. Series D (The Statistician).

CASE STUDY

Taking 1997 base the index numbers of wholesale prices of a commodity are given bellow

Year 1997 1998 1999 2000 2001 2002 2003

Index numbers 100 120 190 200 206 230 300

Construct a new series taking 2000 as base.

Unit V: Time Series Analysis

CHAPTER 10 — Introduction to Time series analysis

Learning Objective:

After studying the unit, students will be able to:

- Know what time series analysis is
- Understand the uses of time series analysis
- Explain the advantages of time series analysis

Structure

10.1 Introduction of Time Series

A time series is a contiguous set of data points that are arranged in chronological order and span a specific time interval. In contrast to cross-sectional data, which represents a single moment in time, this is different.

Investment professionals use time series analysis to track the changes in specific data points, such as the value of a financial asset, over a defined period of time. The data points are gathered at regular intervals. There are no explicit temporal restrictions or boundaries that must be followed. This suggests that the data can be collected in a way that offers the essential information to the investor or analyst who is studying the activity.

A time series refers to any variable that exhibits variations over time. When engaging in investment activities, a time series is frequently employed to monitor the fluctuations in the price of a financial instrument over a period of time. This can be monitored over a brief duration, such as the hourly price fluctuations of a securities during a typical workday, or over an extended duration, such as the monthly closing prices of a security over a span of five years.

Time series analysis is a technique that enables us to examine the fluctuations in an asset, security, or economic variable over a specific timeframe. Furthermore, it can be used to evaluate the relationship between changes in the chosen data point and adjustments in other variables during the same timeframe.

Time series are widely utilised in many non-financial contexts, such as monitoring the temporal evolution of population dynamics. Put simply, a time series refers to a collection of data that is arranged chronologically based on when it occurred. It refers to the sequence of events. In this scenario, time serves as a means to identify significant points of reference for the entire matter. Time can be quantified using units such as hours, days, months, or years. A time series depicts the relationship between two variables. Time is one of these variables, while every other quantity that can be measured numerically is the other. The link between a variable and time does not necessarily exhibit a positive correlation in terms of its rate of change. The relationship does not invariably deteriorate. The direction of movement may vary for individuals at different periods. Can you identify a comparable scenario? One

example of such data is the recorded temperature in a specific city over a given week or month.

10.2. Uses of Time Series

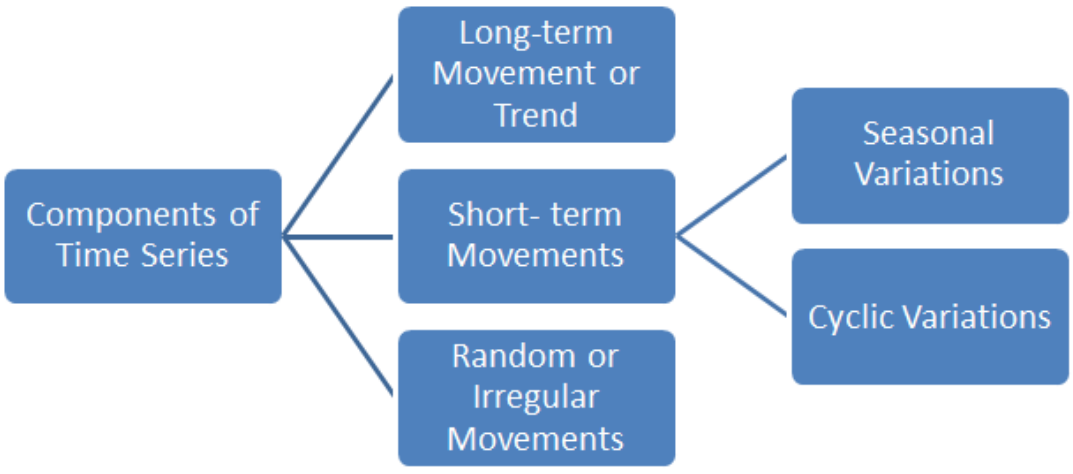
- The most important reason to study time series is that it helps us predict how a variable will act in the future based on what it has done in the past.
- It helps with business planning because it lets you compare how well the business is doing now with how well it should be doing.
- From time series, we can look at how the phenomenon or variable in question has behaved in the past.
- We can look at how the values of different variables change at different times, places, etc. and compare them.

10.3. Components for Time Series

A time series is made up of its different pans, which are the reasons or forces that change the value of an observation over time. There are four types of parts that make up a time series.

- Trend
- Seasonal Variations
- Cyclic Variations
- Random or Irregular movements

Changes that happen every year or every few years are called seasonal and cyclic variations.



10.4 Trends

The trend indicates the long-term direction of the data, whether it has been increasing or decreasing. A trend refers to a consistent and gradual movement over a significant period of time, typically represented by an average value. The notion that the growth or drop must consistently occur in the same direction within a specific time frame is not universally valid.

The trends can exhibit upward, downward, or stagnant movement at various time intervals. The industrial and factory count, and educational institutions are all illustrative of variables that undergo changes over time.

Linear Trend and Non-Linear Trend

If we plot the values of the time series on a graph based on time t. The type of trend can be seen in how the data groups together. If the data tend to fall in a straight line, the trend is linear. If not, the trend is not linear (Curvilinear).

It's a change that will last longer. In this case, we look at the number of observations we have and decide for ourselves what is long term. It shows that a time series moves in the same direction in a fairly smooth, steady, and slow way. Think about how the weather changes to understand what "long term" means.

Over an extended duration, such as 50 years, these variables may undergo periodic fluctuations. If there were only 20 years of data available, this long-term evolution would appear as a discernible pattern. Nevertheless, the availability of many hundred years of data allows for the observation of significant long-term changes. These movements exhibit a regular pattern as they gradually and consistently increase or decrease in the same direction. The trend may be linear or non-linear (curvilinear). Here are some instances of secular trends:

- Rise in prices,
- Rise in environmental contamination,
- A rise in the demand for wheat,
- rise in the literacy rate,

The decline in mortality resulting from scientific advancements. An uncomplicated method for identifying a pattern in seasonal data is to examine the means over a specific duration. The fluctuations in the averages over time indicate the presence of a discernible pattern in the

provided series. There exist more formal methods to determine whether a time series has a trend.

10.5 Summary

- A time series is a sequential collection of data points that occur chronologically over a period of time. Contrary to cross-sectional data, which captures a singular moment in time, this is distinct.
- A time series is employed in investment to monitor the fluctuations of specific data points, such as the value of a financial asset, over a specific duration. The data points are collected at consistent intervals. There are no specific time constraints or limitations that need to be adhered to. This implies that the data can be gathered in a manner that provides the necessary information to the investor or analyst examining the activity.
- A time series refers to any variable that exhibits variations over time. When making investments, a time series is commonly employed to monitor the fluctuations in the price of a security over a period of time. This can be monitored over a brief duration, such as the hourly price fluctuations of a securities during a typical workday, or over an extended duration, such as the monthly closing prices of a security over a span of five years.

10.6 Self-Assessment Questions

1. What are the 4 components of time series?
2. What is a time series?
3. What are the types of time series?
4. What j time series and methods of time series?
5. What is the importance of time series forecasting
6. Discuss the characteristics of time series?
7. Discuss the two models of time series?
8. Discuss the assumptions of time series?
9. Mention the limitations of time series?
10. Discuss the objectives of time series analysis?

10.7 References / Reference Reading

- Hand, D.J. (2004). Measurement theory and practice: The world through quantification. London: Arnold.
- Mann, Prem S. (1995). Introductory Statistics (2nd ed.). Wiley.
- Upton, G., Cook, I. (2008) Oxford Dictionary of Statistics, OUP.

CHAPTER 11 — Types of Time series

Leaning Objective:

After studying the unit, students will be able to:

- Know what time series methods are there
- Understand the meaning of periodic fluctuations
- Explain the difference between periodic and non-periodic fluctuations

Structure

11.1 Periodic Fluctuations

Some parts of a time series tend to happen over and over again over a certain amount of time. They do things in a regular way.

Seasonal Variations

These are the recurrent forces that consistently operate for a period of less than one year. Throughout the year, individuals partake in a monotonous or almost identical regimen on a daily basis. The shift will be seen in a time series if the data is collected at regular intervals such as hourly, daily, weekly, quarterly, or monthly.

These differences emerge either from natural factors or as a consequence of regulations implemented by persons. Seasonal changes generally result from the unique seasons or weather conditions. For example, crop production varies in response to the changing seasons, whereas the demand for umbrellas and raincoats rises during wet weather. Likewise, the demand for electric fans and air conditioners surges throughout the summer season. Artificial customs such as celebrations, conferences, routines, trends, and occasions like matrimony have noticeable effects. They happen in a recurring manner, year after year. An upsurge in a specific season should not be construed as a sign that the firm is progressing.

11.2 Cyclic Variations

Cyclic changes refer to recurring patterns in a time series that occur over a span of more than one year. This oscillation occurs over a duration exceeding one year. A cycle refers to a complete duration of time. The phrase "Business Cycle" is occasionally employed to characterize this trend.

The concept consists of four components: periods of prosperity, periods of adversity, and periods of improvement. The cyclic changes might either be regular or irregular. The fluctuations in business performance are contingent upon the interplay and synergy of many economic variables.

11.3 Movements

The observed deviationin researched variable is also influenced by another reason. They do not vary in a predictable manner; rather, they change randomly. These changes are abrupt, unpredictable, and uncontrollable. These factors encompass seismic activities, armed conflicts, inundations, food shortages, and several other calamities.

After removing the trend and cyclical fluctuations from a set of time series data, the remaining data may or may not exhibit randomness. Various approaches to analyses this type of series involve investigating if "random fluctuations can be accounted for using probability models such as moving average or autoregressive models." This implies that we can determine whether there is any remaining cyclical fluctuation in the residuals. Residual variation, often known as incidental or unpredictable fluctuations, refers to these unexplained variations that occur without any apparent cause. For instance, the price of steel could increase due to several factors such as a factory strike, a brake malfunction accident, a flood, an earthquake, or a war.

11.4 “Mathematical Model for Time Series Analysis”

Mathematically, a time series is given as

yt = f (t)

Here, vt is the value of the variable under study at time t. If the population is the variable under study at the various time periods t 1, t2, t3, tn. Then the time series is

t: t 1, t2, t3, ,tn

yt: yt1 , yt2, yt3, ytn

or t: tl, t2, t3, ,tn

yt: yl, y2, y3, ,yn

11.5 Additive and Multiplicative Models Trend Analysis

Analysis of a time series is breaking a time series into its different parts so that each part can be studied on its own. To analyze a time series, you need to separate and measure its different parts. When we look at a series of events over time, we try to answer the following questions.

- What would be the variable's value at different time points if it were solely influenced by long-term movements?
- What alterations transpire in the value of the variable as a result of seasonal fluctuations?
- How much and in what way has the variable been influenced by cyclical fluctuations?
- The influence of irregular fluctuations has been assessed.

Most of the time, you need to study a time series to make estimates and predictions. A good forecast should be based on predictions of the different kinds of changes. Trend, seasonal, and cyclical changes should all be predicted in their own ways. When there are irregular movements, these predictions become less likely. So, it is important to separate and measure the different kinds of changes in a time series.

A value of a time series variable that is thought of as the sum of the effects of its parts. Either the multiplicative or the additive model is used to describe the parts of a time series.

Let Y = original observation, T = trend component, S =seasonal component, C =cyclical component, and I =irregular component.

11.6 Summary

- It is clear that the trends can go up, down, or stay the same at different points in time. But the big picture must be going up, down, or staying the same. Its population, agricultural output, manufactured goods, number of births and deaths, number of industries or factories, and number of schools or colleges are all examples of things that change over time.

- If we plot the values of the time series on a graph based on time t. The type of trend can be seen in how the data groups together. If the data tend to fall in a straight line, the trend is linear. If not, the trend is not linear (Curvilinear).

- It's a change that will last longer. In this case, we look at the number of observations we have and decide for ourselves what is long term. It shows that a time series moves in the same direction in a fairly smooth, steady, and slow way. Think about how the weather changes to understand what "long term" means.

11.7 References/References Readings

- Mann, Prem S. (1995). Introductory Statistics (2nd ed.). Wiley.
- Upton, G., Cook, I. (2008) Oxford Dictionary of Statistics, OUP.

CHAPTER 12 — Methods of Construction of Seasonal Indices

Learning Objective:

After studying the unit, students \sill be able to:

- Know what seasonal indices are
- Understand the methods of construction for seasonal indices
- Explain the difference between each method

Structure

12.1 Simple Average

Methods of Constructing Seasonal Indices

Seasonal indices can be made in four different ways.

1. Simple averages method
2. Ratio to trend method
3. Percentage moving average method
4. Link relatives' method

This is the easiest and simplest way to study how the seasons change. This method is used when the only parts of the time series variable are the seasonal and random parts. When you take the average of data from the same time period (let's say the first quarter of each year), you get rid of the random component and are left with only the seasonal component. The seasonal indices are then made from these averages. It has to do with the following:

If data is provided on a monthly basis:

- Calculate the monthly average of the raw data for each year.
- Add up all the numbers that have to do with a month. It means to add up all the January values for each year. Do the same thing for each month.
- Find the average of the monthly numbers. To do this, divide the total amount for each month by the number of years. For example, if data for the last five years is available on a monthly basis, there will be five numbers for January. To get the average for January, you have to add up all of these numbers and divide by five. Get these numbers for every month. They may be X1, X2, X3..... X12.
- Calculate the mean of the monthly averages by dividing the sum of the averages by 12, represented as X1+X2+X3+...+X1212=X
- Calculate the sum of the monthly averages by using 100 as the average for each month. The percentage for the average of January (X1) can be calculated by dividing the average of January by the average of monthly averages and multiplying the result by 100.
- The expression "X1X×100" can be simplified as "100 times X1X". If the total amounts of each month were utilised instead of the mean values, the outcome would remain same.

Merits and Demerits

This is the easiest way to measure how the seasons change. But this method is based on an unrealistic idea that the data doesn't show the trend or cyclical changes.

12.2 Ratio to Trend Method

This method is used when there are no cyclical changes in the data, which means that the time series variable Y is made up of trend, seasonal, and random parts. We can write Y = T.S.R with symbols. Among the steps that go into figuring out seasonal indices are:

1. Use the method of least squares to find the trend values for each month, quarter, etc.

2. Divide the original numbers by the trend numbers that go with them. This would get rid of the data's trend values.

3. The quotients are multiplied by 100 to get the percentage.

Merits and Demerits

It is an objective way to measure how the seasons change. But it is very hard to understand and doesn't work if there are cyclical changes.

12.3 Link Relatives' Method

This method presupposes that the trend is characterized by linearity and that cyclical fluctuations adhere to a consistent pattern. The linked relatives display the extent of deviation between the current period (quarter or month) and the preceding period, expressed as percentages. By determining the correlation between relatives and calculating their mean value, the impact of random and cyclical components is diminished. The trend is also addressed through the process of modifying chain relatives.

This method for calculating seasonal indices involves the following steps:

1. Find the Link Relative (L.R.) for each period by dividing that period's number by the number from the period before it. For example, Link relative of 3rd quarter=figure of 3rd quarter ' figure of 2nd quarter x 100.
2. Find the average of the number of link relatives in a given quarter (or month) from different years. This can be done with either A.M. or Md. Theoretically, the second one is better because the first one gives too much weight to extreme things.
3. These averages are turned into chained relatives by assuming that the chained relative of the first quarter (or month) is 100. Chained relative (C.R.) for the current period (quarter or month) = C.R. of the previous period L.R. of the current period 1 100.

4. Use the last quarter to figure out the C.R. for the first quarter (or month) (or month). This is given by C.R. of the last quarter (month) x average L.R. of the first quarter (month) 100 a. In general, this value is different from 100 because of how the data has changed over time. The chained relatives that were found above need to be changed to account for this trend. The adjustment factor

d=14new C.R for 1st quarter-100forquarterly data

d=112new C.A.R for 1st month-100for monthly data

On the assumption that the trend is linear d, 2d, 3d, etc, is respectively subtracted from the 2nd, 3 rd , 4th , etc quarter (or month).

1. To get seasonal indices, write the adjusted chained relatives as a percentage of their average.
2. Make sure that the total of these indices is 400 for quarterly data and 1200 for monthly data.

Merits and Demerits

The ratio to moving average and the ratio to trend methods are more complicated than this method. But this method is based on the idea that trends move in a straight line, which may not always be the case.

12.4 Summary

- The trend shows whether the data have been going up or down over a long period of time. A trend is an overall, long-term, average, smooth movement. It's not always true that the increase or decrease has to go in the same direction over a certain time period.
- It is clear that the trends can go up, down, or stay the same at different points in time. But the big picture must be going up, down, or staying the same. Its population, agricultural output, manufactured goods, number of births and deaths, number of industries or factories, and number of schools or colleges are all examples of things that change over time.
- If we plot the values of the time series on a graph based on time t. The type of trend can be seen in how the data groups together. If the data tend to fall in a straight line, the trend is linear. If not, the trend is not linear (Curvilinear).
- It's a change that will last longer. In this case, we look at the number of observations we have and decide for ourselves what is long term. It shows that a time series moves in the same direction in a fairly smooth, steady, and slow way. Think about how the weather changes to understand what "long term" means.

- Man-made traditions such as festivals, conventions, habits, styles, and events like marriage have discernible impacts. They occur repeatedly, year after year. An increase in a particular season should not be interpreted as an indication that the business is improving.
- The change in the variable being studied is also caused by another thing. They don't change in a regular way; instead, they change at random. These changes are sudden, can't be predicted, and can't be controlled. These forces include earthquakes, wars, floods, famines, and any other disasters.
- Most of the time, you need to study a time series to make estimates and predictions. A good forecast should be based on predictions of the different kinds of changes. Trend, seasonal, and cyclical changes should all be predicted in their own ways. When there are irregular movements, these predictions become less likely. So, it is important to separate and measure the different kinds of changes in a time series.
- A value of a time series variable that is thought of as the sum of the effects of its parts. Either the multiplicative or the additive model is used to describe the parts of a time series.
- The moving averages method is one way to find the underlying trend in a set of data by smoothing out the highs and lows. Estimating the trend can also be done with other tools, such as regression analysis.

12.5 Self-Assessment Questions

1. Define an average method?
2. What is the link relativity methods?
3. What is an additive model in statistics?
4. Why multiplicative model is the most commonly used model in time series analysis?
5. What are the two models in time series analysis?
6. What is the best method for seasonal variation?
7. How do you calculate the average seasonal effect?
8. What do you mean by seasonal variation?
9. How do you calculate seasonal indices?
10. How do you calculate seasonal indices by ratio to trend method?

12.6 References/ Reference Readings

- Hand, D.J. (2004). Measurement theory and practice: The world through quantification. London: Arnold.
- Mann, Prem S. (1995). Introductory Statistics (2nd ed.). Wiley.
- Upton, G., Cook, I. (2008) Oxford Dictionary of Statistics, OUP.

CASE STUDY

The following data relate to the income of the people and the General index number ofprices of a certain region. Calculate:

1) Real income and
2) Index numbers of real income with 1996

1996	800	100
1997	819	105
1998	825	110
1999	876	120
2000	920	125
2001	938	140
2002	924	140