Sahana Prasad

Advanced Statistical Methods

Springer

Advanced Statistical Methods

Sahana Prasad

Advanced Statistical Methods

Springer

Sahana Prasad
R V University
Bengaluru, Karnataka, India

ISBN 978-981-99-7256-2 ISBN 978-981-99-7257-9 (eBook)
https://doi.org/10.1007/978-981-99-7257-9

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Paper in this product is recyclable.

Advanced Statistical Methods

MR STAT

Hi, I am Mr STAT and I believe in lucid learning. In this book, I have simplified all the complex concepts into manageable chunks of theory in a narrative manner. I have also added plenty of interesting charts, meaningful graphs, thoughtful tables and also a few data comics!

All these will help you have a clear understanding of statistical concepts with an analytical bent of mind!

Simplistic

Theories with

Accurate

Trendy graphs

MISS TICS

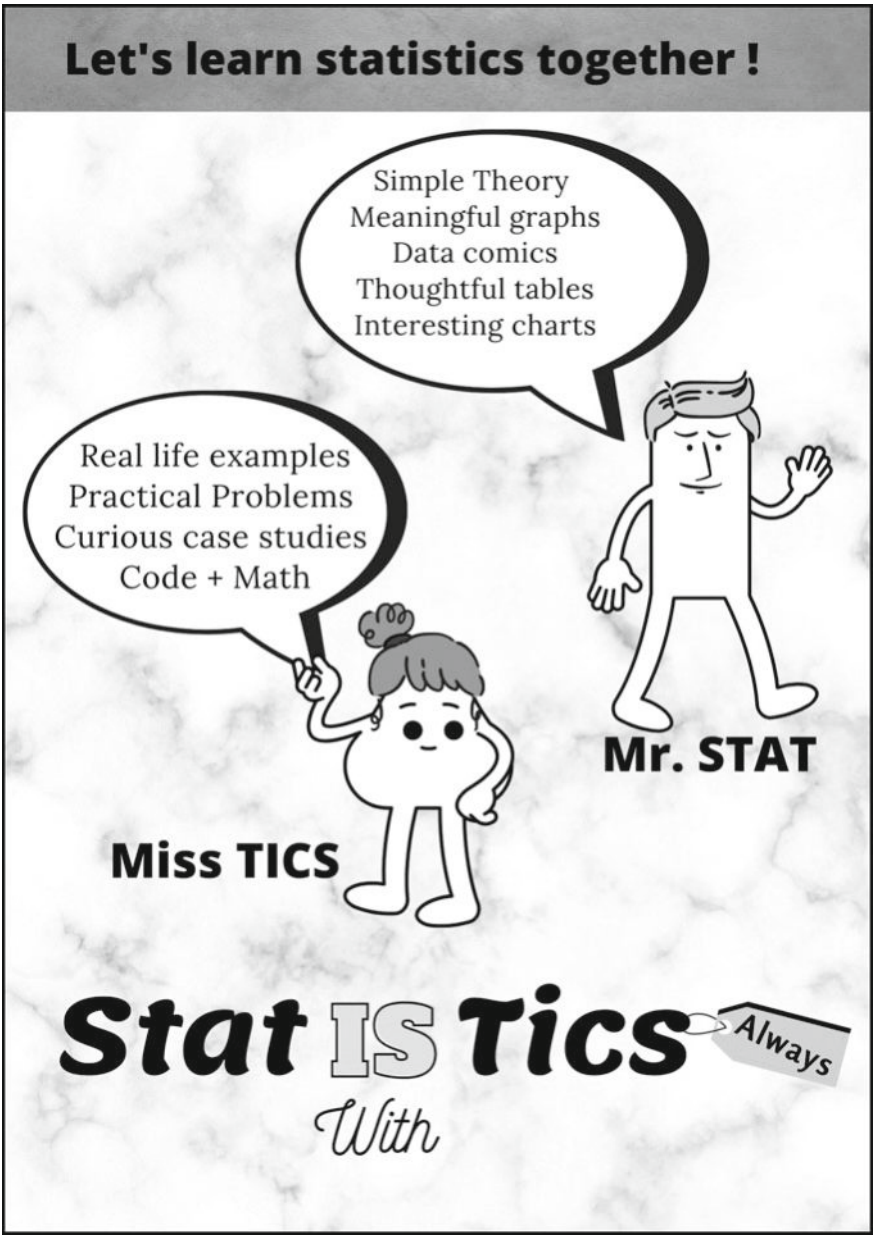
Hi, I am Miss TICS and I always emphasise learning with examples and encourage the application of these concepts in reality. Let's code, and decode the inner meaning of data together. Let's hunt for relevant programming tools to solve the mystery of numbers. With this, you can get hands-on experience in solving every problem with a statistical solution, meaningful interpretation and a reliable conclusion. Let's get Sherlocked with curious case studies!

Thoughtful

Interesting examples with

Case

Studies


Let's learn statistics together !
Simple Theory
Meaningful graphs
Data comics
Thoughtful tables
Interesting charts
Real life examples
Practical Problems
Curious case studies
Code + Math
Mr. STAT
Miss TICS
Stat IS Tics
Always
With

Acknowledgments

I would like to thank the Almighty who has been instrumental in guiding me in envisaging and creating this book. His unseen hand has led me up the path of writing a useful and simple book.

I thank Springer Nature and the Senior Editor Shamim Ahmed in particular, who encouraged and supported me in this arduous journey. I am grateful to them as this book could not see the light of the day, otherwise.

My heartfelt thanks to my husband Mr. L. V. Prasad for his support and encouragement. My daughter Sharanya, son Amit Vikram, my son-in-law Vineeth Patil and most importantly, my dear grandson Rihaan who have stood by me and lent their suggestions and time in shaping this book. I am also thankful to my father B. N. Vishwanath who has always encouraged me in all my endeavors.

Ms. Srishma Sunku, my collaborator, has helped me immensely by proofreading, creating infographics, adding new content and editing the book thoroughly. I am indebted to her as the book has been improved widely by her efforts. She is the editor who has verified the contents as well as added the infographics, wherever needed.

My students, colleagues, fellow researchers and other coworkers whose suggestions and requirements have dictated the content of this book need special mention. I thank each and every one of them profusely.

About This Book

Welcome to "Advanced Statistical Methods." This book is designed to be a comprehensive and practical guide to understanding and utilizing some of the fundamental tools in statistical analysis for real-world applications. Whether you are a student, researcher, or professional in any field, this book aims to equip you with the necessary knowledge and skills to harness the power of data and make informed decisions.

The world we live in today is inundated with data, and the ability to analyze and interpret this information is crucial. In this context, regression analysis serves as a powerful tool to establish relationships between variables and predict future outcomes. With regression, we can explore cause-and-effect relationships and identify patterns that can help us better understand and optimize various processes. Through real-world case studies, we will explore how regression can be applied to fields as diverse as economics, marketing, healthcare, and more.

Index numbers, another vital topic covered in this book, play an integral role in measuring changes in economic and social indicators over time. Understanding how to construct and interpret index numbers is essential in tracking inflation, calculating purchasing power, and comparing economic performance across different periods. We will examine various methodologies and case studies that demonstrate the practical significance of index numbers in different scenarios.

Vital statistics offer a unique perspective on understanding population dynamics and health trends. We will delve into the analysis of birth rates, death rates, population growth, and other demographic indicators that have profound implications for public policy, healthcare planning, and social development. Through the lens of real-life applications, we will witness how vital statistics can inform decision-making at local, national, and global levels.

Time series analysis is another critical aspect of this book, enabling us to discern patterns and trends within data that evolves over time. From financial market forecasting to climate change predictions, time series analysis is central to numerous fields. We will explore various time series models, including moving averages, auto regressive models, and more, while examining case studies that demonstrate their efficacy and limitations.

Throughout this book, the focus remains firmly on practical applications. Each chapter includes hands-on examples and case studies that bridge the gap between theory and real-world problem-solving. We will leverage popular statistical software and programming languages to demonstrate how these methodologies can be implemented effectively.

It is my sincere hope that this book will empower you to harness the full potential of statistical tools in your professional endeavors. By the end of this journey, you should feel confident in applying these methods to analyze data, draw meaningful insights, and make informed decisions that drive positive change.

So, let's embark on this enriching journey through the realm of applied analytics together, as we explore the diverse landscape of regression, index numbers, vital statistics, and time series, and unlock the potential of data-driven decision-making.

Happy learning and analyzing!

Contents

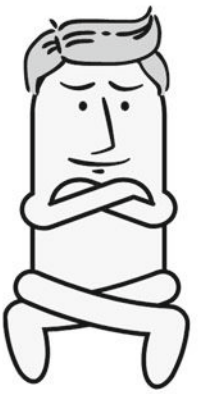
List of Figures

List of Tables

Chapter 1
Regression

S. Prasad, *Advanced Statistical Methods*, https://doi.org/10.1007/978-981-99-7257-9_1

REGRESSION



Mr STAT

- Introduction and basic
- terminologies.
- Types of regression - Linear, quadratic, multiple regression,
- logistic regression.
- Methods of regression.
- Methods and steps to regression analysis.

Miss TICS

- Case studies and applications of regression in various fields of study like: sales forecast, credit risk assessment, marketing, .
- When to use which regression method.
- Caution while using regression analysis.

Fig. 1.1 A comic on reactive and predictive data

What is regression analysis?

Regression analysis is a statistical technique used to examine the relationship between a dependent variable and one or more independent variables. It aims to find the best-fitting mathematical model that describes the relationship between variables and enables predictions or explanations based on that model.

Why use regression analysis?

Regression analysis is used for several purposes, including:

Prediction: It helps predict the value of the dependent variable based on known values of independent variables. (Copyright 2015 by Modern Analyst Media L.L.C) (Fig. 1.1).

Explanation: It provides insights into how independent variables are related to the dependent variable, allowing researchers to understand the underlying mechanisms and factors influencing the outcome.

Control: Regression analysis can help control for the effects of confounding variables by including them as independent variables in the model.

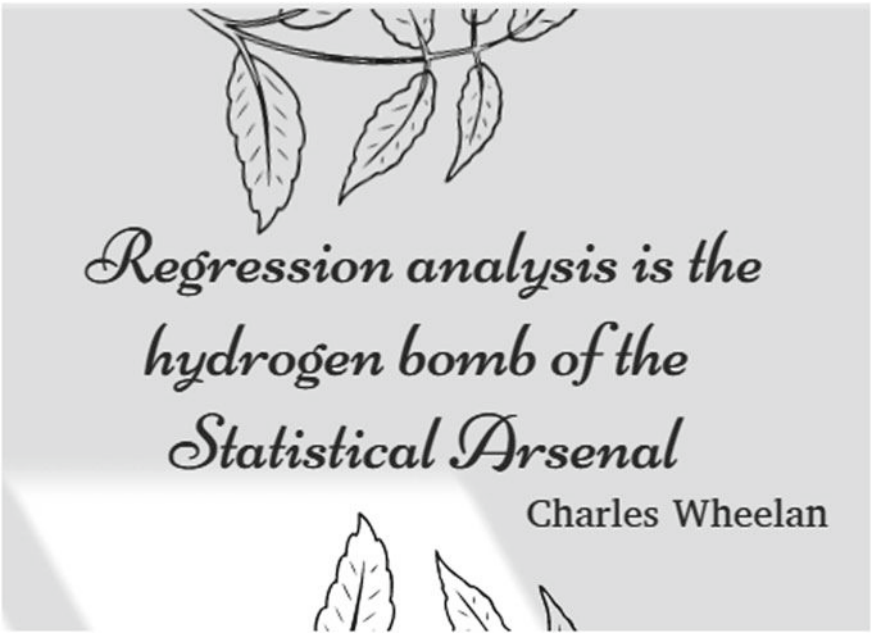


Fig. 1.2 Quote on regression

How does regression analysis work?

Regression analysis involves several steps:

Data collection: Gather data on the dependent variable and independent variables of interest.

Model selection: Choose the appropriate regression model based on the nature of the data and research question (e.g., linear regression, multiple regression, logistic regression, etc.).

Model estimation: Estimate the coefficients of the regression equation that best fit the data.

Model assessment: Evaluate the goodness of fit of the model to determine how well it explains the relationship between variables.

Interpretation: Interpret the coefficients of the regression equation to understand the direction and magnitude of the relationship between variables.

Where is regression analysis used?

Regression analysis is widely used in various fields, including (Fig. 1.2):

Economics: To study the relationship between economic variables such as GDP, inflation, and unemployment.

Finance: To predict stock prices, analyze risk factors, and understand the determinants of asset returns.

Social sciences: To examine the impact of variables such as education, income, and demographics on social outcomes.

Marketing: To analyze the relationship between marketing expenditures, customer behavior, and sales.

Healthcare: To identify risk factors, predict patient outcomes, and analyze the effectiveness of treatments.

When to use regression analysis?

Regression analysis can be used when there is a relationship between a dependent variable and one or more independent variables.

The relationship can be reasonably assumed to be linear or can be transformed to approximate linearity.

Sufficient data is available to estimate the model parameters reliably.

Note: Regression analysis assumes certain assumptions about the data, such as linearity, independence of errors, and homoscedasticity, which should be evaluated before applying the method.

1.1 Difference Between Correlation and Regression

See Table 1.1.

Table 1.1 Differences between correlation and regression

Correlation	Regression
Relationship between variables in terms of numbers	Mathematical relationship between variables
Examines the relationship and there are no distinction b/w independent and dependent variables	The difference between variables is clearly highlighted
Spurious correlations can exist	No spurious correlations
We understand how the variables move together	We study the changes in one variable (dependent) when there is a unit change in another (independent) variable

1.2 Different Types of Regression Analysis

Regression analysis encompasses various types of regression techniques, each designed to address specific research questions or data characteristics. Here are some of the main types of regression and their typical usage:

Linear Regression: Linear regression is the most fundamental type of regression analysis. It examines the linear relationship between a dependent variable and one or more independent variables. It is used when the relationship between variables can be approximated by a straight line. Linear regression finds applications in various fields, including economics, social sciences, finance, and engineering.

Multiple Regression: Multiple regression extends linear regression by considering the relationship between a dependent variable and multiple independent variables. It is used to analyze how a combination of factors influences the dependent variable. Multiple regression is valuable when studying complex relationships and determining the relative importance of various predictors.

Polynomial Regression: Polynomial regression is an extension of linear regression that allows for modeling nonlinear relationships between variables. It involves using higher-order polynomials to fit curves instead of straight lines. Polynomial regression is used when the relationship between variables exhibits a curvilinear pattern.

Logistic Regression: Logistic regression is utilized when the dependent variable is categorical or binary (e.g., yes/no, success/failure). It models the probability of an event occurring based on independent variables. Logistic regression finds applications in areas such as medical research, social sciences, and marketing, where predicting binary outcomes is important.

Ridge Regression and Lasso Regression: Ridge regression and Lasso regression are used when dealing with multicollinearity, a situation where independent variables are highly correlated. These techniques introduce regularization to the regression model, helping to reduce the impact of multicollinearity and prevent overfitting. Ridge regression and Lasso regression are particularly useful in situations where there are more predictors than observations.

Time Series Regression: Time series regression is employed when analyzing data collected over time, where the order of observations is crucial. It considers the temporal component and the relationship between the dependent variable and one or more independent variables. Time series regression is commonly used in financial analysis, economics, and forecasting.

Nonlinear Regression: Nonlinear regression is used when the relationship between the dependent variable and independent variables cannot be adequately described by a linear equation. It allows for more flexible modeling of complex relationships by using nonlinear functions or transformations. Nonlinear regression finds applications in various fields, such as physics, biology, and environmental sciences.

These are some of the main types of regression analysis. Choosing the appropriate regression technique depends on the nature of the data, the research question, and the underlying assumptions. It is essential to understand the characteristics of each regression type to select the most suitable approach for a given analysis.

1.3 Regression Lines

At the heart of regression analysis lies the regression line, a fundamental concept that enables us to model and predict values based on observed data. By fitting a line through observed data points, regression analysis provides valuable insights into the strength and direction of the relationship. Understanding regression lines is crucial for researchers, analysts, and decision-makers across various disciplines, enabling them to make informed predictions and draw meaningful conclusions from data.

1.3.1 Definition

A regression line is a straight line that best represents the relationship between two variables in a scatter plot. It serves as a mathematical model to estimate the average value of one variable (dependent variable) based on the known values of another variable (independent variable).

1.3.2 Types of Regression Lines

There are two primary types of regression lines (Fig. 1.3):

a. Simple Linear Regression Line: This type of regression line represents the relationship between two variables when there is a linear association. It is defined by the equation

$$Y = a + bX,$$

where Y is the dependent variable, X is the independent variable, a is the intercept, and b is the slope.

b. Multiple Regression Line: In cases where more than one independent variable influences the dependent variable, multiple regression is used. The regression line becomes a multidimensional plane or hyperplane that represents the relationship among all the variables involved.

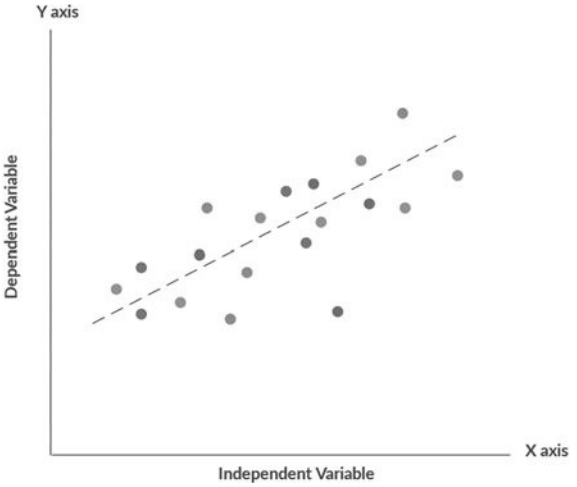


Fig. 1.3 Regression lines

1.3.3 Fitting a Regression Line

To find the best-fitting regression line, the method of least squares is commonly employed. This method minimizes the sum of the squared differences between the observed data points and the corresponding predicted values on the line. The resulting regression line passes through the center of the data cloud, capturing the general trend of the relationship.

1.3.4 Interpreting a Regression Line

The regression line provides valuable insights into the relationship between variables. Key elements of interpretation include:

a. **Slope**: The slope of the regression line (b) represents the change in the dependent variable associated with a one-unit change in the independent variable. A positive slope indicates a positive relationship, while a negative slope suggests a negative relationship.
b. **Intercept**: The intercept (a) represents the predicted value of the dependent variable when the independent variable is zero. It is the point where the regression line intersects the y-axis.
c. **Goodness of Fit**: The closeness of the data points to the regression line is a measure of the goodness of fit. The coefficient of determination (R-squared) quantifies the proportion of the variation in the dependent variable that is explained by the independent variable(s).

1.3.5 Practical Applications

Regression lines have a wide range of applications across various fields:

a. **Economics**: Regression analysis is used to study the relationships between economic variables such as GDP, inflation, and unemployment.
b. **Social Sciences**: Regression lines help analyze the impact of independent variables on dependent variables, such as studying the effect of education level on income.
c. **Finance**: Regression analysis is used to model stock prices, analyze risk factors, and estimate returns on investments.
d. **Medicine**: Regression lines aid in predicting patient outcomes based on medical variables, such as predicting the progression of a disease based on various biomarkers.

1.3.6 Properties of Regression Lines

Regression lines can be considered as graphical representations of regression equations. Since there are two regression equations there will be two regression lines that will follow the properties mentioned below:

The lines intersect at a point, which is denoted as $(\overline{X}, \overline{Y})$.

The closer the lines are to each other, the greater the correlation will be.

When the correlation is equal to 0, there is no correlation or linear relationship between the regression lines. Therefore, the lines will be perpendicular to each other.

When correlation is equal to 1, the regression lines are perfectly correlated. The two lines will coincide and there will be one single line visible.

1.4 Regression Equations

The regression equation is a fundamental component of regression analysis as it provides a mathematical representation of the relationship between variables. Regression equations provide a mathematical representation of the relationship between variables in a regression model. They enable researchers and analysts to quantify and predict the impact of independent variables on the dependent variable. Understanding regression equations is crucial for conducting robust statistical analyses and making informed decisions in various domains.

1.4.1 Definition

A regression equation is a mathematical formula that describes the relationship between variables in a regression model. It represents the best-fitting line or curve that minimizes the difference between the observed data points and the predicted values.

The equation is typically written in the form: $Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n$, where Y is the dependent variable, X_1, X_2, …, Xn are the independent variables, b_0 is the intercept, and b_1, b_2, …, b_n are the regression coefficients.

1.4.2 Components of a Regression Equation

a. Dependent Variable (Y): The variable being predicted or explained by the regression model.
b. Independent Variables (X_1, X_2, …, X_n): The variables used to predict or explain the dependent variable.
c. Intercept (b_0): The value of the dependent variable when all independent variables are zero.
d. Regression Coefficients (b_1, b_2, …, b_n): The coefficients that represent the change in the dependent variable associated with a one-unit change in the corresponding independent variable, holding other variables constant.

1.4.3 Interpretation of Regression Coefficients

Each regression coefficient in the equation has its own interpretation. For instance, a positive coefficient indicates that an increase in the corresponding independent variable leads to an increase in the dependent variable, while a negative coefficient suggests the opposite. The magnitude of the coefficient reflects the strength of the relationship between the variables.

1.4.4 Different Regression Equations

Simple Linear Regression Equation: In simple linear regression, there is only one independent variable. The equation takes the form: $Y = b_0 + b_1X$, where Y is the dependent variable, X is the independent variable, b_0 is the intercept, and b1 is the slope coefficient. The slope coefficient represents the change in Y associated with a one-unit change in X.

Multiple Regression Equation: Multiple regression involves more than one independent variable. The equation expands as follows: $Y = b_0 + b_1X_1 + b_2X_2 + \cdots$

$+b_n X_n$, where Y is the dependent variable, X_1, X_2, …, X_n are the independent variables, and b_1, b_2, …, b_n are the respective regression coefficients. Each coefficient represents the change in Y associated with a one-unit change in the corresponding independent variable, holding other variables constant.

1.4.5 Practical Applications

Regression equations have widespread applications across various fields. They are extensively used in economics, finance, social sciences, healthcare, and many other disciplines. For example, they can be employed to predict sales based on advertising expenditure, determine the impact of education on income, or forecast stock prices using historical data. Regression provides valuable insights into relationships between variables, helping professionals make evidence-based decisions and predictions.

Some examples are as follows:

Economics

Demand and price elasticity: Regression analysis can be used to estimate the impact of price changes on the demand for a product or service, helping businesses optimize pricing strategies.

Economic growth: Researchers often employ regression analysis to study the determinants of economic growth, examining factors such as investment, education, and technological progress.

Social Sciences

Education and income: Regression analysis can explore the relationship between education level and income, controlling for other variables such as age, gender, and occupation, to assess the impact of education on earnings.

Crime rates: Regression analysis can be used to investigate the factors influencing crime rates, such as poverty, unemployment, and law enforcement expenditures.

Finance: Stock price prediction: Regression models can be employed to analyze historical stock prices and other financial indicators to forecast future stock prices, aiding investors in making informed decisions.

Risk analysis: Regression analysis can help financial institutions assess the risk associated with lending by examining factors such as credit score, income, and debt-to-income ratio.

Marketing: Customer behavior: Regression analysis can be used to understand consumer behavior, such as the impact of advertising expenditure on sales or the influence of pricing on customer purchasing decisions.

Market segmentation: Regression models can assist in identifying different customer segments based on demographic or psychographic variables, enabling targeted marketing campaigns.

Healthcare: Disease prognosis: Regression analysis can be applied to predict the prognosis of a disease based on patient characteristics, laboratory results, and medical history, aiding in treatment planning.

Health outcomes: Researchers often use regression models to investigate the relationship between health outcomes (e.g., mortality rates, patient satisfaction) and various factors, including treatment methods and patient demographics.

1.5 Simple Linear Regression

It is a statistical approach that helps in understanding the relationship between 2 continuous variables that are quantitative. It is used in all those fields where there is quantitative data like forecasting, ML models, etc. The important thing in forecasting simple linear regression is that the dependent variable must be continuous and the independent variable can be measured as continuous or categorical.

This method has 2 objectives:

(a) To model the relationship between variables under study.
(b) Forecasting new observations based on relationship derived.

In simple linear regression there are 2 variables—independent (X) and dependent (Y) and we have to find a relationship between X and Y and we can have the following approaches:

Deterministic: We can predict the output variables using a function of independent variable.

For example, all well-known formulae such as area of cylinder, volume of a triangle, Ohm's law, etc.

Random: There may be no relationship between variables or relationships may exist only during certain time period and not during other periods. Most of the studies defines a random relationship as a "No relationship between variables."

For example, we spend Rs.500 on a lucky draw and win up to Rs.100. In this scenario, we definitely cannot say that for every 500 rupees we spend, we can earn a 100 rupee.

1.5.1 Properties of Regression Coefficients in Simple Linear Regression

In simple linear regression, there is only one independent variable, and the regression equation takes the form:

$$Y = b_0 + b_1 X$$

where Y is the dependent variable, X is the independent variable, b_0 is the intercept, and b_1 is the regression coefficient.

The regression coefficient, b_1, represents the slope of the regression line, which indicates the change in the dependent variable (Y) associated with a one-unit change in the independent variable (X). It quantifies the strength and direction of the linear relationship between the variables.

1.5.2 Interpreting the Regression Coefficient

Positive coefficient ($b_1 > 0$): A positive coefficient indicates a positive relationship between the independent and dependent variables. For every one-unit increase in the independent variable, the dependent variable is expected to increase by the value of the coefficient. Conversely, for a one-unit decrease in the independent variable, the dependent variable is expected to decrease by the coefficient value.

Negative coefficient ($b_1 < 0$): A negative coefficient indicates a negative relationship between the independent and dependent variables. For every one-unit increase in the independent variable, the dependent variable is expected to decrease by the absolute value of the coefficient. Similarly, for a one-unit decrease in the independent variable, the dependent variable is expected to increase by the absolute value of the coefficient.

Zero coefficient ($b_1 = 0$): A coefficient of zero suggests no linear relationship between the independent and dependent variables. In this case, changes in the independent variable have no effect on the dependent variable.

It's important to note that the regression coefficient provides information about the average relationship between the variables in the sample data. However, it does not necessarily imply causation, and other factors should be considered to establish causal relationships.

To estimate the regression coefficients (b_0 and b_1), statistical techniques such as the method of least squares are commonly used to find the best-fitting line that minimizes the sum of squared differences between the observed data points and the predicted values.

Understanding the regression coefficients in simple linear regression is crucial for interpreting and drawing insights from the relationship between the variables under study.

The two regression coefficients have the following properties

- Both the coefficients must have the same sign.
- If one of the regression coefficients is greater than 1 then the other should be less than 1.

The geometric mean of regression coefficients is equal to correlation coefficient,

$$r = \mp \sqrt{b_{yx} * b_{xy}}$$

1.5.3 Estimating the Regression Equations from the Given Data in Case of Simple Linear Regression

To estimate the regression equation in simple linear regression, you need a set of data consisting of pairs of observations for the dependent variable (Y) and the independent variable (X). The regression equation represents the relationship between these two variables.

Let's assume you have a dataset with the following observations:

$$(X_1, Y_1)\ (X_2, Y_2)\ (X_3, Y_3) \ldots (X_n, Y_n)$$

To estimate the regression equation, you need to find the slope (β_1) and the intercept (β_0) that best fit the data. The regression equation is given by:

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

To estimate the slope (β_1) and intercept (β_0), you can use the least squares method. This method minimizes the sum of the squared differences between the observed Y

values and the predicted Y values based on the regression equation. The formulas to estimate the slope and intercept are as follows:

$$\beta_1 = \Sigma\left((X_i - \overline{X})(Y_i - \overline{Y})\right) / \Sigma\left((X_i - \overline{X})^2\right)$$

$$\beta_0 = \overline{Y} - \beta_1 * \overline{X}$$

where

$\mathbf{X_i}$ is the value of the independent variable for observation i.

$\bar{\mathbf{X}}$ is the mean of the independent variable values.

$\mathbf{Y_i}$ is the value of the dependent variable for observation i.

$\bar{\mathbf{Y}}$ is the mean of the dependent variable values.

By calculating the slope (β_1) and intercept (β_0) using these formulas, you can estimate the regression equation for the data.

1.5.4 Examples of Calculating Regression Equations from Given Data

Example 1: Housing Prices

Suppose you want to estimate the relationship between the size of houses (in square feet) and their corresponding prices (in dollars). You collect data for 10 houses, as follows (Table 1.2).

To estimate the regression equation, we'll calculate the slope ($\beta 1$) and intercept ($\beta 0$) using the formulas mentioned earlier:

Step 1: Calculate the means: $\overline{X} = (1500 + 1800 + \cdots + 2000 + 1600) / 10 = 1850$

Table 1.2 Table of prizes of houses based on their sizes

House size (X)	Price (Y)
1500	2,00,000
1800	2,50,000
2200	3,00,000
1400	1,90,000
2400	3,20,000
1700	2,30,000
1900	2,60,000
2100	2,80,000
2000	2,70,000
1600	2,10,000

$$\overline{Y} = (200{,}000 + 250{,}000 + \cdot + 270{,}000 + 210{,}000 \quad)/10 = 240{,}000$$

Step 2: Calculate $\frac{\sum((Xi-\overline{X})(Yi-\overline{Y}))}{\sum((Xi-\overline{X})^2)}$

$$= \frac{((1500-1850)(200{,}000-240{,}000)+(1800-1850)(250{,}000-240{,}000) + \cdot + (1600-1850)(210{,}000-240{,}000))}{((1500-1850)^2+(1800-1850)^2+\cdot+(1600-1850)^2)}$$
$$= \frac{-35{,}000}{1{,}650{,}000} \approx -0.0212$$

Step 3: Calculate β_0:

$$\beta_0 = \overline{Y} - \beta_1 * \overline{X} = 240{,}000 - (-0.0212)(1850) \approx 278{,}900$$

The estimated regression equation is: price = 278,900 − 0.0212 * Size of the house

Example 2: Advertising and Sales: Suppose you want to examine the relationship between advertising expenditures (in dollars) and corresponding sales (in units) for a certain product. You collect data for 8 different advertising campaigns, as follows (Table 1.3).

Following the same steps as in the previous example:

Step 1: Calculate the mean:

$$\overline{X} = (1000 + 1200 + \cdot + 1300 + 1400 \quad)/8 = 1250$$
$$\overline{Y} = (50 + 60 + \cdot + 65 + 70 \quad)/8 = 62.5$$

Step 2: Calculate

Table 1.3 Advertising expenditure tabulated against the sales

Advertising expenditure (X)	Sales (Y)
1000	50
1200	60
1500	70
900	45
1800	80
1100	55
1300	65
1400	70

Table 1.4 Production tabulated with respect to the electricity consumed

Production (X)	Electricity usage (Y)
4.51	2.48
3.58	2.26
4.31	2.47
5.06	2.77
5.64	2.99
4.99	3.05
5.29	3.18
5.83	3.46

$$\frac{\sum((\mathrm{Xi}-\overline{\mathrm{X}})(\mathrm{Yi}-\overline{\mathrm{Y}}))}{\sum((\mathrm{Xi}-\overline{\mathrm{X}})^2)}$$

$$= \frac{\begin{array}{c}(1000-1250)(50-62.5)+(1200-1250)(60-62.5)\\ +\cdot\cdot+(1400-1250)(70-62.5)\end{array}}{(1000-1250)^2+(1200-1250)^2+\cdot\cdot+(1400-1250)^2)}$$

$$= \frac{-8750}{350{,}000} \approx -0.025$$

Step 3: Calculate $\beta_0 \beta_0 = \overline{\mathrm{Y}} - \beta_1 * \mathrm{X} = 62.5 - (-0.025) * (1250) \approx \mathbf{93.75}$.

Step 4: The estimated regression equation is: **Sales = 93.75 − 0.025 * Advertising expenditure**.

Example: For the electricity usage and production data given below compute the regression equations. In this dataset production is the independent variable, and electricity usage Y is the dependent variable (Table 1.4).

Solution:

See Table 1.5.

The equation of regression can be written as,

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{x}$$

$$\beta_1 = \frac{\mathrm{n}\sum \mathrm{xy} - (\sum \mathrm{x} \sum \mathrm{y})}{\mathrm{n}\sum \mathrm{x}^2 - (\sum \mathrm{x})^2}$$

$$\beta_0 = \overline{\mathrm{y}} - \beta_1 \overline{\mathrm{x}}$$

$$\beta_1 = \frac{12(169.25) - (58.62)(34.15)}{12(291.22) - (58.62)^2}$$

$$\beta_1 = \frac{2031 - 2001.873}{3494.64 - 3436.3044}$$

Table 1.5 Calculation of regression equation for the electricity consumed and production values

Production (X)	Electricity usage (Y)	XY	X^2
4.51	2.48	11.1848	20.34
3.58	2.26	8.0908	12.82
4.31	2.47	10.6457	18.58
5.06	2.77	14.0162	25.60
5.64	2.99	16.8636	31.81
4.99	3.05	15.2195	24.90
5.29	3.18	16.8222	27.98
5.83	3.46	20.1718	33.98
4.7	3.03	14.241	22.09
5.61	3.26	18.2886	31.4721
4.9	2.67	13.083	24.01
4.2	2.53	10.626	17.64
58.62	**34.15**	**169.626**	**291.22**

$$= \frac{29.127}{58.3356}$$

$$\beta_1 = 0.4993$$

$$\overline{x} = \frac{\sum x}{n} = \frac{58.62}{12} = 4.885 \quad \overline{y} = \frac{\sum y}{n} = \frac{34.15}{12} = 2.8458$$

$$\beta_0 = \overline{y} - \beta_1 \overline{x}$$

$$= 2.8458 - 0.4993(4.885)$$

$$= 2.8458 - 2.4391$$

$$\beta_0 = 0.4067$$

$$y = 0.4067 + 0.4993x$$

Example 3. A survey was conducted to study the relationship between expenditure (in Rs.) on accommodation (x) and expenditure on food and entertainment (y) and the following results were obtained (Table 1.6).

Coefficient of correlation, r = 0.57.

Write down the regression equation and estimate the expenditure on food and entertainment if the expenditure on accommodation is Rs.200.

Table 1.6 Table to study the relationship between the expenditure and accommodation

Item	Mean	Standard deviation
Expenditure on accommodation	173	63.15
Expenditure on food and entertainment	47.8	22.98

Solution:
Let expenditure on accommodation be X.
Let expenditure on food and entertainment be Y.
Given: $\overline{x} = 173 \quad \overline{y} = 47.8 \quad \sigma_x = 63.15 \quad \sigma_y = 22.98 \quad r = 0.57$.
Given X =200, find Y.
We have to use Y on X

$$Y - \overline{y} = byx\left(X - \overline{X}\right)$$

$$b_{yx} = \frac{r\sigma_y}{\sigma_x} = 0.57\left(\frac{22.98}{63.15}\right) = 0.2074$$

$$Y - 47.8 = 0.2074 \quad (200 - 173)$$

$$Y - 47.8 = 5.5998$$

$$\mathbf{Y = 53.3998}$$

Note: The method of estimating the regression equation in simple linear regression is closely related to finding the line of best fit. In fact, the regression equation itself represents the line of best fit.

The line of best fit, also known as the fitted line or the regression line, is a straight line that represents the best approximation of the relationship between the independent variable (X) and the dependent variable (Y) based on the given data points. It is called the "best fit" because it minimizes the overall distance between the line and the data points.

In simple linear regression, the line of best fit is determined by estimating the slope (β1) and the intercept (β0) of the regression equation. The slope represents the rate of change of the dependent variable with respect to the independent variable, while the intercept represents the value of the dependent variable when the independent variable is zero.

The process of estimating the regression equation involves finding the values of β0 and β1 that minimize the sum of the squared differences between the observed Y values and the predicted Y values based on the regression equation. This method is known as the least squares method.

In summary, the method of estimating the regression equation in simple linear regression and finding the line of best fit are closely related. The regression equation provides the mathematical representation of the line of best fit, while the least squares method is used to determine the values of the slope and intercept that minimize the overall error between the line and the data points.

1.5.5 Line of Best Fit

In regression analysis, the line of best fit refers to the straight line that best represents the relationship between the independent variable(s) and the dependent variable in a

linear regression model. It is also known as the regression line or the least squares line.

The line of best fit is determined by minimizing the sum of the squared differences between the observed data points and the corresponding predicted values on the line. This technique is called ordinary least squares (OLS) regression.

Mathematically, the line of best fit is represented by the equation:

$$y = mx + b$$

where y is the dependent variable (the variable we want to predict).

x is the independent variable (the variable used to predict y).

m is the slope of the line, representing the change in y for a one-unit change in x.

b is the y-intercept, representing the value of y when x is zero.

The slope (m) and the y-intercept (b) are estimated using statistical methods to minimize the sum of squared residuals, which are the differences between the observed y-values and the predicted y-values on the line. The line that minimizes the sum of squared residuals is considered the best fit line.

Once the line of best fit is determined, it can be used to predict the values of the dependent variable for given values of the independent variable(s) within the range of the data. The line can also be used to assess the strength and direction of the relationship between the variables, as well as to identify any outliers or influential points that deviate significantly from the line.

Example 1: Height and Weight

Let's consider a dataset that includes the heights (in inches) and weights (in pounds) of individuals. We want to find the line of best fit to predict weight based on height (Table 1.7).

Step 1: Calculate the means of x and y:

Table 1.7 Table showing the height and weight of individuals

Height (Y) (inches)	Weight (X) (pounds)	$x - \overline{x}$	$y - \overline{y}$	$(x - \overline{x})^2$	$(y - \overline{y})^2$	$(x - \overline{x}) * (y - \overline{y})$
125	64	−3	−11.7	9	136.89	−34.9
140	68	1	3.3	1	10.89	3.3
160	72	5	23.3	25	542.89	116.5
110	60	−7	−26.7	49	713.89	186.9
135	66	−1	−1.7	1	2.89	1.7
155	70	3	18.3	9	335.89	55
825	**400**	**−2**	**4.8**	**94**	**1743.34**	**330.5**

Table 1.8 Table with distance covered along with time taken

Distance (X)	Time (Y)	$x - \overline{x}$	$y - \overline{y}$	$(x - \overline{x})^2$	$(y - \overline{y})^2$	$(x - \overline{x}) * (y - \overline{y})$
50	3.5	−125	−6.73	15,625	45.15	845.63
100	6.8	−75	−3.43	5625	11.79	257.25
150	9.2	−25	−1.03	625	1.06	25.75
200	12.1	25	1.87	625	3.5	46.75
250	14.8	75	4.57	5625	20.89	343.13
300	17.5	125	7.27	15,625	52.91	1139.38

$$\text{mean(x)} = \overline{x} = (64 + 68 + 72 + 60 + 66 + 70)/6 = 67\text{kg}$$
$$\text{mean(y)} = \overline{y} = (125 + 140 + 160 + 110 + 135 + 155)/6 = 136.7\,\text{inches}$$

Step 2: Calculate the differences from the means: x −mean(x): [−3, 1, 5, −7, −1, 3] y −mean(y): [−11.7, 3.3, 23.3, −26.7, −1.7, 18.3].

Step 3: Calculate the squared differences: $\sum(x - \overline{x})^2 = [9, 1, 25, 49, 1, 9]$:

$$\sum(y - y)^2 = [136.89, 10.89, 542.89, 713.89, 2.89, 335.89]$$

Step 4: Calculate the product of the differences: [−34.9, 3.3, 116.5, 186.9, 1.7, 55].

Step 5: Calculate the slope (m): $m = \frac{\Sigma[(x - \text{mean}(x))(y - \text{mean}(y))]}{\Sigma[(x - \text{mean}(x))^2]} = 330.5/94 \approx 3.511$.

Step 6: Calculate the y-intercept (b): $b = \overline{y} - (m * \overline{x}) = 136.7 - 3.511 * 67 \approx -95.24$.

Therefore, the line of best fit is: $y = 3.511x - 95.24$.

Example 2: Time and Distance.

Let's consider a dataset that records the time (in hours) it takes to travel a certain distance (in miles). We want to find the line of best fit to predict time based on distance (Table 1.8).

Step 1: Calculate the means of x and y:

$$\text{Mean (x)} = (50 + 100 + 150 + 200 + 250 + 300)/6 = 175\,\text{miles}$$
$$\text{Mean (y)} = (3.5 + 6.8 + 9.2 + 12.1 + 14.8 + 17.5)/6 = 10.23\,\text{h}$$

Step 2: Calculate the differences from the means. (Column 3 & 4).

Step 3: Calculate the squared differences: (Column 5 & 6).

Step 4: Calculate the product of the differences: (Column 7 & 8).

Table 1.9 Table scores in sports and pre board exam of few students

Scores in sports (X)	Scores in preboard exam (Y)
10	20
12	10
15	18
10	12

Step 5: Calculate the slope (m):

$$m = \frac{\Sigma\left[(x - \text{mean}\ (x))(y - \text{mean}\ (y))\right]}{\Sigma\left[(x - \text{mean}\ (x))^2\right]} = \frac{3253.89}{44450} \approx 0.0733$$

Step 6: Calculate the y-intercept (b):

$$b = \overline{y} - m * \overline{x} = 10.23 - 0.0733 * 175 \approx -0.42$$

Therefore, the line of best fit is: $y = 0.0733x - 0.42$

1. Consider the scores of selected students in college with respect to sports and exams. Find the regression equation and line of best fit (Table 1.9).

Solution:

See Table 1.10.

$$\overline{x} = \frac{\sum x}{n} = \frac{47}{4} = 11.75 \quad \overline{y} = \frac{\sum y}{n} = \frac{60}{4} = 15$$
$$a = \frac{\sum xy}{\sum x^2} = \frac{5}{16.75} = 0.2985$$
$$b = \frac{\sum y}{n} = 15$$
$$y = 0.2985x + 15.$$

Table 1.10 Solution table for calculating the regression equation for the scores obtained in sports and preboard exam

x	y	$X = x - \overline{x}$	$Y = y - \overline{y}$	XY	X^2
10	20	−1.75	5	−8.75	3.0625
12	10	0.25	−5	−1.25	0.0625
15	18	3.25	3	9.75	10.5625
10	12	−1.75	−3	5.25	3.0625
47	60	0	0	5	16.75

Note: This method can also be done without transforming x and y. We compute all the required values and use the normal equations.

Compute and fit multiple regression model to the following data (Table 1.11).

Calculate the regression sums:

$$\sum x_1^2 = \Sigma X_1^2 - \left[\frac{(\Sigma X_1)^2}{n}\right] = 38767 - \left[\frac{(555)^2}{8}\right] = 38767 - \left[\frac{308025}{8}\right]$$

$$\sum x_1^2 = 38767 - 38503.125 = 263.875$$

$$\sum x_2^2 = \Sigma X_2^2 - \left[\frac{(\Sigma X_2)^2}{n}\right] = 2823 - \left[\frac{(145)^2}{8}\right]$$

$$= 2823 - 2628.125 = 194.875$$

$$\sum x_2^2 = 194.875$$

$$\Sigma x_1 y = \Sigma X_1 y - \frac{\Sigma X_1 \Sigma y}{n}$$

$$= 101895 - \frac{(555)(1452)}{8}$$

$$= 101895 - \frac{805860}{8}$$

$$= 101895 - 100732.5$$

$$\Sigma x_1 y = 1162.5$$

$$\Sigma x_2 y = \Sigma X_2 y - \frac{\Sigma X_2 \Sigma y}{n}$$

$$= 25364 - \frac{(145)(1452)}{8}$$

$$= 25364 - 26317.5$$

$$\Sigma x_2 y = -953.5$$

Table 1.11 Table of values to compute the regression lines

y	X_1	X_2	$X_1{}^2$	$X_2{}^2$	X_1 y	X_2 y	X_1X_2
140	60	22	3600	484	8400	3080	1320
155	62	25	3844	625	9610	3875	1550
159	67	24	4489	576	10,653	3816	1608
179	70	20	4900	400	12,530	3580	1400
192	71	15	5041	225	13,632	2880	1065
200	72	14	5184	196	14,400	2800	1008
212	75	14	5625	196	15,900	2968	1050

$$\begin{aligned}\Sigma x_1 x_2 &= \sum X_1 X_2 - \left(\frac{\Sigma X_1 \Sigma X_2}{n}\right)\\ &= 9859 - \frac{(555)(145)}{8}\\ &= 9859 - 10059.375\\ \Sigma x_1 x_2 &= -200.375.\end{aligned}$$

We calculate b_0, b_1 etc.

$$\begin{aligned}b_0 &= \overline{y} - b_1 \overline{X}_1 - b_2 \overline{X}_2\\ b_1 &= \frac{\sum x_2^2 \left(\sum x_1 y\right) - \left(\sum x_1 x_2\right)\left(\sum x_2 y\right)}{\sum x_1^2 \left(\sum x_2^2\right) - \left(\sum x_1 x_2\right)^2}\\ b_2 &= \frac{\sum x_1^2 \left(\sum x_2 y\right) - \left(\sum x_1 x_2\right)\left(\sum x_1 y\right)}{\sum x_1^2 \left(\sum x_2^2\right) - \left(\sum x_1 x_2\right)^2}\end{aligned}$$

$$\begin{aligned}b_1 &= \frac{194.875 \times 1162.5 - (-200.375)(-953.5)}{263.875(194.875) - (-200.375)^2}\\ &= \frac{226542.1875 - 191057.5625}{51422.64063 - 40150.14063} = \frac{35484.625}{11272.5}\\ \mathbf{b_1} &= \mathbf{3.1479}\end{aligned}$$

$$\begin{aligned}b_2 &= \frac{263.875 \times (-953.5) - (-200.375)(1162.5)}{263.875(194.875) - (-200.375)^2}\\ &= \frac{-251604.8125 + 232935.9375}{11272.5} = \frac{-18668.875}{11272.5}\\ &= \mathbf{-1.6557}\end{aligned}$$

$$\begin{aligned}\overline{y} &= \sum \frac{y}{n} = \frac{1452}{8} = 181.5\\ \overline{X}_1 &= \sum \frac{X_1}{n} = \frac{555}{8} = 69.375\\ \overline{X}_2 &= \sum \frac{X_2}{n} = \frac{145}{8} = 18.125\\ b_0 &= \overline{y} - b_1 \overline{X}_1 - b_2 \overline{X}_2\\ &= 181.5 - (3.1479)(69.375) - (-1.6557)(18.125)\\ &= 181.5 - 218.3856 + 30.0096 = -6.867\\ \hat{y} &= b_0 + b_1 x_1 + b_2\\ x_2 &= -6.867 + 3.148 x_1 - 1.656 x_2.\end{aligned}$$

1.5.6 Evaluation of Model Goodness of Fit in Simple Regression: Understanding R-squared and Adjusted R-squared

Introduction: In simple regression analysis, assessing the goodness of fit of a model is crucial to understand how well it explains the relationship between the dependent variable and the independent variable. Two commonly used measures for evaluating model fit are R-squared (coefficient of determination) and adjusted R-squared. This article aims to explain these measures, their interpretation, and their significance in assessing the quality of a simple regression model.

R-squared: R-squared is a statistical measure that represents the proportion of the variance in the dependent variable that can be explained by the independent variable(s) in a regression model. It ranges between 0 and 1, where 0 indicates that the independent variable(s) has no explanatory power, and 1 indicates a perfect fit.

Interpretation of R-squared: The interpretation of R-squared depends on the context and the nature of the data. A higher R-squared value suggests that a larger proportion of the variation in the dependent variable can be accounted for by the independent variable(s). Conversely, a lower R-squared value indicates that the independent variable(s) has less explanatory power. It is important to note that R-squared alone does not determine the accuracy or validity of a model, and other factors should be considered for a comprehensive evaluation.

Limitations of R-squared: While R-squared is a widely used metric, it has certain limitations. R-squared tends to increase when additional independent variables are added to the model, even if those variables do not have any meaningful impact on the dependent variable. This is where adjusted R-squared comes into play.

Adjusted R-squared: Adjusted R-squared addresses the limitation of R-squared by adjusting for the number of independent variables in the model. It takes into account the degrees of freedom and penalizes the inclusion of irrelevant variables. Adjusted R-squared is always lower than or equal to R-squared.

Interpretation of Adjusted R-squared: Adjusted R-squared is a more conservative measure of model fit compared to R-squared. It considers the trade-off between the number of variables and the improvement in model fit. A higher adjusted R-squared indicates that the independent variable(s) explain a larger proportion of the variation in the dependent variable, considering the number of variables in the model. It helps prevent overfitting by penalizing the inclusion of unnecessary variables.

Comparing R-squared and Adjusted R-squared: When deciding between R-squared and adjusted R-squared, it is essential to consider the complexity of the model and the number of variables included. If the model contains only one independent variable, R-squared and adjusted R-squared will be identical. However, as the number of variables increases, adjusted R-squared becomes a more reliable measure of model fit.

Conclusion: Evaluating the goodness of fit is crucial in simple regression analysis to determine the effectiveness of the model in explaining the relationship between the dependent and independent variables. R-squared provides a measure of the proportion of variance explained, while adjusted R-squared adjusts for the number of variables in the model. Both measures have their significance and should be interpreted in conjunction with other evaluation criteria to draw valid conclusions about the model's quality.

1.6 Interpretation of the Standard Error of the Estimate

The standard error of the estimate (SE) is a measure of the average distance between the observed values of the dependent variable and the predicted values obtained from a regression model. It quantifies the variability or scatter of the data points around the regression line. A smaller SE indicates less dispersion of the data points and a better fit of the model.

Interpreting the SE involves considering the context of the data and the scale of the dependent variable. Typically, the SE is reported in the same units as the dependent variable. For example, if the dependent variable represents sales in dollars, the SE would be expressed in dollars.

The SE can be used to estimate the precision of the predicted values. It provides a range within which we can expect the actual values of the dependent variable to fall with a certain level of confidence. Specifically, approximately 68% of the observed values are expected to lie within one standard error of the estimate from the predicted values, and about 95% are expected to lie within two standard errors.

Example 1: Suppose you conducted a survey to estimate the average height of adults in a particular city. After collecting data from a sample of 100 individuals, you calculated the mean height to be 170 cm with a standard error of 2 cm. In this case, the standard error indicates the average amount of sampling variation in the mean height estimates. You can interpret it as follows: "Based on our sample, we estimate that the average height of adults in this city is 170 cm, with a margin of error of 2 cm."

Example 2: Let's say you performed an experiment to investigate the effect of a new drug on blood pressure. After conducting the experiment on a sample of 50 participants, you calculated the mean reduction in blood pressure to be 10 mmHg, with a standard error of 1.5 mmHg. In this scenario, the standard error represents the precision of the estimate. You could interpret it as: "We found that the average reduction in blood pressure due to the new drug is 10 mmHg, and this estimate is reasonably precise with a standard error of 1.5 mmHg."

Example 3: Consider a regression analysis where you're examining the relationship between income (independent variable) and education level (dependent variable). After analyzing a dataset of 200 individuals, you obtain a regression coefficient of

0.25 with a standard error of 0.05. In this case, the standard error is associated with the precision of the regression coefficient estimate. You could interpret it as: "The estimated coefficient for income is 0.25, indicating that a one-unit increase in income is associated with a 0.25 unit increase in education level, on average. The estimate is quite precise, given the standard error of 0.05."

Remember, the standard error provides information about the precision or variability of an estimate. It helps quantify the margin of error and indicates the reliability of the estimated parameter.

1.7 Significance of the Model

Assessing the significance of a regression model involves evaluating whether the relationship between the independent variable(s) and the dependent variable is statistically significant. This is typically done by conducting hypothesis tests, such as the t-test or F-test, on the regression coefficients.

The t-test is used to test the significance of individual coefficients in simple regression or multiple regression models. It determines whether the coefficient is significantly different from zero, implying that the independent variable has a statistically significant impact on the dependent variable. A significant coefficient suggests that the variable is likely to have a meaningful effect on the outcome.

The F-test is used to test the overall significance of the regression model. It assesses whether the regression model as a whole provides a better fit to the data compared to the null model (i.e., a model with no independent variables). A significant F-test indicates that the independent variable(s) collectively have a significant impact on the dependent variable.

In both tests, the significance level (alpha) is specified to determine the threshold for accepting or rejecting the null hypothesis. The commonly used threshold is 0.05, meaning that if the p-value associated with the test statistic is below 0.05, the result is considered statistically significant.

Interpreting the significance of the model involves considering both the statistical and practical significance. A statistically significant model implies that the relationship between the variables is unlikely to have occurred by chance. However, it is important to assess the practical significance by considering the magnitude and direction of the coefficients and their relevance in the specific context or domain.

The standard error of the estimate provides a measure of the average distance between observed and predicted values, while the significance of the model determines whether the relationship between the variables is statistically meaningful. Both interpretations are essential in evaluating the accuracy and practical relevance of the regression model.

1.8 A Few Case Studies that Demonstrate the Application of Linear Regression

Housing Prices: A common use case for linear regression is predicting housing prices. A dataset can be collected, including features such as the size of the house, number of bedrooms, location, and age. By applying linear regression, you can estimate the relationship between these features and the house prices, allowing you to make predictions for new houses based on their characteristics.

Sales Forecasting: Linear regression can be used to predict sales based on various factors, such as advertising expenditure, promotional activities, and historical sales data. By analyzing the relationship between these variables, you can develop a model that forecasts future sales based on the given inputs.

Stock Market Analysis: Linear regression can be applied to analyse stock prices and predict future trends. By examining historical stock data and considering relevant factors like company performance, economic indicators, or news sentiment, you can build a regression model to estimate future stock prices.

Customer Lifetime Value (CLV): Linear regression can be used to predict customer lifetime value, which is the projected revenue a company expects to earn from a customer during their entire relationship. By using historical data on customer purchases, engagement, and other relevant factors, you can create a model to estimate the future value of customers and optimize marketing strategies accordingly.

Demand Forecasting: Linear regression can help predict the demand for a product or service. By analyzing historical sales data and considering variables such as price, promotional activities, seasonality, and economic indicators, you can develop a regression model that estimates future demand and assists in inventory planning and production optimization.

These case studies demonstrate how linear regression can be applied in various domains to make predictions and gain insights from data. Keep in mind that real-world applications often involve more complex models and techniques, but linear regression serves as a foundational tool in data analysis and prediction.

1.9 Polynomial Regression

In the realm of predictive modeling, polynomial regression stands out as a versatile technique that can capture complex relationships between variables. While linear regression assumes a linear connection between the dependent and independent variables, polynomial regression extends this concept by introducing higher-order terms. By incorporating polynomial functions, this regression method can uncover nonlinear patterns and provide a more accurate representation of the underlying data.

1.9.1 Understanding Polynomial Regression

Polynomial regression is a form of regression analysis where the relationship between the independent variable (X) and the dependent variable (Y) is modeled as an nth-degree polynomial. In simple terms, it allows us to fit a curve instead of a straight line to the data points. This technique is particularly useful when a linear relationship fails to capture the true nature of the data.

Polynomial Regression Equation:

The polynomial regression equation can be represented as:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \cdots + \beta_n X^n + \varepsilon$$

Here, Y is the dependent variable, X represents the independent variable, βn denotes the regression coefficients for the respective powers of X, and εis the error term. The degree of the polynomial is denoted by "n," which determines the flexibility of the curve.

1.9.2 Benefits of Polynomial Regression

Capturing Nonlinear Relationships: Polynomial regression enables the discovery of nonlinear relationships that may exist in the data. By including higher-order terms, it can represent curves, bends, and fluctuations, providing a better fit to complex patterns.

Improved Model Accuracy: The ability to capture nonlinear relationships often results in improved prediction accuracy compared to linear regression. Polynomial regression allows for a more precise representation of the data, especially when the underlying relationship is curved or has multiple turning points.

Flexibility and Interpretability: Polynomial regression offers flexibility in model selection. By choosing an appropriate degree for the polynomial, researchers can control the trade-off between bias and variance. Additionally, the coefficients of the polynomial equation provide insights into the magnitude and direction of the relationships.

Extrapolation Capability: Polynomial regression can extend predictions beyond the range of the observed data. While extrapolation should be approached cautiously, polynomial models can be useful for estimating values outside the known range based on the fitted curve.

1.9.3 Challenges and Considerations

Overfitting: Polynomial regression models with high degrees can become overly complex and prone to overfitting. Overfitting occurs when the model captures noise or random variations in the data, leading to poor generalization on new, unseen data. Regularization techniques such as ridge regression or Lasso regression can help mitigate this issue.

Model Selection: Choosing the appropriate degree of the polynomial is crucial. A degree that is too low may result in an oversimplified model that fails to capture the underlying patterns, while a degree that is too high may lead to overfitting. Model evaluation techniques, such as cross-validation, can aid in determining the optimal degree.

Data Availability: Polynomial regression often requires a sufficient amount of data to accurately estimate the coefficients. As the degree increases, the number of required data points grows exponentially. Insufficient data can lead to unstable and unreliable coefficient estimates.

Polynomial regression provides a powerful tool for modeling nonlinear relationships in predictive analytics. By capturing the complexity of data through curves and bends, it allows for improved accuracy and a deeper understanding of the underlying patterns. While challenges such as overfitting and model selection exist, proper techniques and considerations can maximize the benefits of polynomial regression. When linear regression falls short, polynomial regression emerges as an invaluable technique in the data scientist's toolbox.

Case Study:

Ref: https://online.stat.psu.edu/stat462/node/159/.

78 Blue gills were randomly studied in a lake. The following dataset was obtained.

$\mathbf{X_1}$: Age of fish (Predictor variable).

$\mathbf{X_2}$: length of the fish. (Response variable).

The researchers were interested in how the length of fish is related to age.

Step 1: A scattered plot of data was drawn and suggested a positive trend in data which implies that the age of fish and the length of fish are proportional. However, this trend does not appear to be linear but it shows a curvilinear relationship.

Step 2: Using the data, estimated equation was obtained by:

$$\text{length} = 13.62 + 54.05 * \text{(Age)} - 4.719\ (\text{Age})^2$$

Step 3: Using this equation, a plot was drawn (this is a fitted line plot).

Step 4: By understanding the output of the analysis, we can draw conclusions.

Step 5: When the data was processed in a software, we generated ANOVA table as output. The model summary was $R^2 = 80.11\%$, and confidence interval was (160.386, 171.418).

Step 6: With this output, we can conclude that 80.1% of variation in the length of the fish is affected by the age of the fish. Also, we can be 95% confident that the length of randomly selected fish lies in the confidence interval.

Case Study 2:

Exam Scores and Study Time

Researchers wanted to examine the relationship between study time and exam scores for a group of students. They collected data on the number of hours each student studied (X1) and their corresponding exam scores (X2). The analysis suggested a curvilinear relationship between study time and exam scores.

Step 1: Scatter Plot

A scatter plot was drawn, indicating a curvilinear trend between study time and exam scores.

Step 2: Estimated Equation

An estimated equation was obtained:

$$\text{Exam Score} = 60.72 + 5.28 * (\text{Study Time}) - 0.63 * (\text{Study Time})^2.$$

Step 3: Fitted Line Plot

A plot was drawn using the estimated equation to visualize the fitted relationship between study time and exam scores.

Step 4: Analysis Conclusion

The analysis concluded that the relationship between study time and exam scores is curvilinear.

Step 5: ANOVA Table

An ANOVA table was generated as the output of the analysis.

The model summary showed an R-squared value of 73.89%.

The confidence interval for the predicted exam scores was (82.48, 91.76).

Step 6: Conclusions

Approximately 73.89% of the variation in exam scores can be explained by study time.

With 95% confidence, the exam scores of a randomly selected student are expected to fall within the confidence interval (82.48, 91.76).

Case study 3: Temperature and Plant Growth

Researchers conducted a study to explore the relationship between temperature and plant growth. They recorded the average daily temperature (X_1) and the corresponding plant growth measurements (X_2). The analysis suggested a curvilinear relationship between temperature and plant growth.

Step 1: Scatter Plot

A scatter plot was drawn, indicating a curvilinear trend between temperature and plant growth.

Step 2: Estimated Equation

An estimated equation was obtained:

$$\text{Plant Growth} = 2.34 + 0.85 * \text{(Temperature)} - 0.09 * \text{(Temperature)}^2.$$

Step 3: Fitted Line Plot

A plot was drawn using the estimated equation to visualize the fitted relationship between temperature and plant growth.

Step 4: Analysis Conclusion

The analysis concluded that the relationship between temperature and plant growth is curvilinear.

Step 5: ANOVA Table

An ANOVA table was generated as the output of the analysis.
The model summary showed an R-squared value of 67.52%.
The confidence interval for the predicted plant growth was (6.24, 9.78).

Step 6: Conclusions

Approximately 67.52% of the variation in plant growth can be explained by temperature.

With 95% confidence, the plant growth for a randomly selected day is expected to fall within the confidence interval (6.24, 9.78).

1.10 Multiple Regression

Multiple regression is a statistical technique used to explore the relationship between a dependent variable and two or more independent variables. It extends the principles of simple linear regression to account for multiple predictors and enables us to understand how different factors interact to influence the outcome of interest.

- Dependent Variable: The dependent variable, also known as the outcome variable or the response variable, is the variable that we want to predict, explain, or understand better using multiple regression. It represents the quantity or characteristic that is affected or influenced by the independent variables. For example, in a study examining the factors affecting a person's salary, the dependent variable would be the individual's salary.
- Independent Variables: Independent variables, also known as predictor variables or explanatory variables, are the variables that are hypothesized to have an impact on the dependent variable. These variables are considered the potential causes or factors that may influence the outcome of interest. In the salary example, the independent variables might include years of experience, level of education, job title, and geographic location.
- Assumptions: Multiple regression relies on several assumptions to ensure the validity and reliability of the results. It is important to assess whether these assumptions hold true in the data before drawing conclusions. The key assumptions of multiple regression include:

Linearity: There should be a linear relationship between the independent variables and the dependent variable. This means that the change in the dependent variable is proportional to the change in the independent variables, holding other variables constant.

Independence: The observations should be independent of each other, meaning that there should be no systematic relationship or dependence between the residuals (the differences between the observed and predicted values) of the regression model.

Homoscedasticity: The residuals should have constant variance across all levels of the independent variables. Homoscedasticity implies that the spread of the residuals should not systematically increase or decrease as the predicted values of the dependent variable change.

Absence of Multicollinearity: The independent variables should not be highly correlated with each other. Multicollinearity occurs when there is a strong linear relationship between two or more independent variables, which can make it difficult to separate their individual effects on the dependent variable.

Violations of these assumptions can lead to biased estimates, invalid inferences, and less accurate predictions. Therefore, it is important to assess these assumptions through various diagnostic tools, such as residual plots, scatterplots, correlation matrices, and tests specifically designed for each assumption.

By ensuring that these assumptions are met or appropriately addressed, researchers can have confidence in the validity of the multiple regression analysis and the interpretations drawn from the results.

Case study 1: Consider the following data (Table 1.12):

Following the same steps as in Example 1, we obtain the following results:

$$b_1 \approx 2.4; b_2 \approx 1.8; b_0 \approx 5.4$$

Table 1.12 Table of values to understand computing multiple regression

Observation	X_1	X_2	Y
1	1	3	12
2	2	5	15
3	3	7	18
4	4	9	21
5	5	11	24

Therefore, the multiple regression equation for this example is:

$$Y \approx 5.4 + 2.4X_1 + 1.8X_2$$

These examples demonstrate the manual calculations involved in determining the regression coefficients and the multiple regression equation for a given set of data. However, it is worth noting that in practice, statistical software or spreadsheet programs are typically used to perform these calculations, as they can handle larger datasets more efficiently and provide additional statistical measures and diagnostics.

In multiple regression analysis, there are several statistical measures that can be used to evaluate the goodness of fit of the model. Here are some commonly used measures, including the R-squared and adjusted R-squared, along with examples:

R-squared (coefficient of determination): R-squared is a measure of how well the regression model fits the observed data. It represents the proportion of the variance in the dependent variable that is explained by the independent variables in the model. R-squared ranges from 0 to 1, with higher values indicating a better fit.

Example: Let's say we have a multiple regression model that predicts a student's exam score (dependent variable) based on their study time, sleep hours, and previous exam score (independent variables). If the R-squared value is 0.80, it means that 80% of the variance in the exam scores can be explained by the study time, sleep hours, and previous exam score included in the model.

Adjusted R-squared: Adjusted R-squared is an adjusted version of R-squared that takes into account the number of predictors in the model. It penalizes the inclusion of unnecessary variables and provides a more conservative measure of the model's goodness of fit.

Example: Suppose we have another multiple regression model with five independent variables predicting housing prices. The R-squared value is 0.75, and the model has 10 predictors. The adjusted R-squared value is 0.70, indicating that the inclusion of some predictors may not contribute significantly to the model's fit, and the adjusted R-squared adjusts for this by slightly reducing the R-squared value.

F-test: The F-test assesses the overall significance of the regression model by comparing the fit of the full model (with predictors) to the fit of a reduced model

(without predictors). It evaluates whether the inclusion of independent variables significantly improves the fit of the model.

Example: Let's consider a multiple regression model that predicts a person's income based on their age, education level, and work experience. The F-test result yields a p-value of 0.001, which is below the significance level of 0.05. This indicates that the regression model, with the included predictors, provides a significantly better fit than a model without any predictors.

These measures, such as R-squared, adjusted R-squared, and the F-test, help assess the goodness of fit of a multiple regression model and provide insights into how well the model explains the relationship between the dependent variable and the independent variables.

1.10.1 A Few Case Studies Where Multiple Linear Regression Has Been Applied

Housing Price Prediction: Multiple linear regression can be used to predict housing prices based on various factors such as location, square footage, number of bedrooms and bathrooms, and other relevant features. The dataset would include information about these independent variables and the corresponding sale prices as the dependent variable. By fitting a multiple linear regression model, one can identify the significant predictors and estimate the impact of each variable on the housing prices.

Sales Forecasting: In retail and sales industries, multiple linear regression can be employed to forecast sales based on factors like advertising expenditure, promotional activities, competitor prices, and other market variables. By analyzing historical sales data along with the independent variables, a regression model can be built to predict future sales figures, aiding in demand planning and resource allocation.

Credit Risk Assessment: Multiple linear regression can be utilized in credit risk assessment to predict the likelihood of default for borrowers. By considering various financial and personal variables of the borrowers, such as income, credit history, debt-to-income ratio, and employment status, a regression model can be built to assess the creditworthiness of individuals or businesses. This information can guide lending institutions in making informed decisions about granting loans and setting appropriate interest rates.

Employee Performance Analysis: Multiple linear regression can be applied in the field of human resources to analyse the factors influencing employee performance. Independent variables could include variables such as education level, years of experience, training programs attended, job satisfaction, and other relevant factors. By examining these predictors, the regression model can help identify the significant factors that contribute to employee performance and guide HR policies and practices.

Each case study requires careful selection of variables and data analysis techniques to build an effective regression model that provides valuable insights.

1.11 Logistic Regression

Logistic regression is a widely used statistical technique in the field of machine learning and data analysis. It is particularly effective for binary classification problems, where the goal is to predict whether an observation belongs to one of two classes. This article aims to provide a comprehensive overview of logistic regression, explaining its underlying principles, assumptions, and practical applications.

1.11.1 What is Logistic Regression?

Logistic regression is a supervised learning algorithm that models the relationship between a set of independent variables (features) and a binary dependent variable (target). Unlike linear regression, which predicts continuous values, logistic regression estimates the probability of an observation belonging to a particular class.

1.11.2 Working Principle

The fundamental concept behind logistic regression is the logistic function, also known as the sigmoid function. The sigmoid function maps any real-valued number to a value between 0 and 1, making it suitable for representing probabilities. In logistic regression, the sigmoid function is used to transform the linear combination of input features into a probability score, indicating the likelihood of the observation belonging to the positive class.

1.11.3 Model Training and Optimization

The logistic regression model is trained using a method called maximum likelihood estimation (MLE). The objective is to find the optimal set of coefficients that maximize the likelihood of observing the given data. This process involves iteratively adjusting the coefficients using optimization algorithms such as gradient descent.

1.11.4 Assumptions of Logistic Regression

Logistic regression assumes that the relationship between the features and the log-odds of the target variable is linear. Additionally, it assumes that the observations are independent and that there is little to no multicollinearity among the independent variables. Violations of these assumptions can lead to biased or inefficient estimates.

1.11.5 Evaluation and Interpretation

To evaluate the performance of a logistic regression model, various metrics such as accuracy, precision, recall, and the receiver operating characteristic (ROC) curve are commonly used. Additionally, logistic regression coefficients can be interpreted as the change in the log-odds of the target variable associated with a one-unit change in the corresponding feature, providing insights into the importance and direction of the variables.

1.11.6 Practical Applications

Logistic regression finds extensive applications across multiple domains, including healthcare, finance, marketing, and social sciences. It is used for predicting disease outcomes, credit risk assessment, customer churn analysis, sentiment analysis, and more. Logistic regression's simplicity, interpretability, and ability to handle both categorical and continuous variables make it a valuable tool in the data scientist's toolkit.

Logistic regression is a powerful and widely used statistical technique for binary classification problems. By estimating the probability of an observation belonging to a particular class, logistic regression provides valuable insights and predictions. Understanding its underlying principles, assumptions, and practical applications can empower data analysts and machine learning practitioners to leverage this versatile tool effectively.

1.11.7 Logit Function

The logit function is a key component of logistic regression, and it plays a crucial role in transforming the linear combination of input features into a probability score. The logit function is the inverse of the sigmoid function and is commonly used to model binary outcomes.

Mathematically, the logit function is defined as the natural logarithm ($base_e$) of the odds ratio. Given a probability p, the odds ratio is calculated as p/(1 −p). Taking the logarithm of the odds ratio yields the logit function, denoted as logit(p).

The formula for the logit function is as follows:

$$\text{logit(p)} = \mathbf{log}\left(\frac{\mathbf{p}}{1-\mathbf{p}}\right)$$

The logit function maps the probability values ranging from 0 to 1 to a range from negative infinity to positive infinity. This transformation is important because it allows for a linear relationship between the independent variables and the log-odds of the target variable in logistic regression.

The logit function has several useful properties. First, it transforms probabilities to a scale where the relationship with the predictors can be modeled linearly. Second, it converts multiplicative relationships into additive relationships, which simplifies the estimation process. Lastly, the logit function is symmetric, meaning that it maps both extremely low and extremely high probabilities to values close to negative infinity and positive infinity, respectively.

In logistic regression, the logit function is used to model the relationship between the independent variables and the log-odds of the target variable. By estimating the coefficients of the independent variables through model training, logistic regression calculates the log-odds of the target variable for a given set of predictor values. These log-odds can then be transformed back into probabilities using the sigmoid function, allowing for classification into one of the two classes.

Overall, the logit function is a fundamental mathematical tool in logistic regression, enabling the modeling and prediction of binary outcomes based on the relationship between input features and the log-odds of the target variable.

1.11.8 Binary Outcome

In logistic regression, a binary outcome variable is a categorical variable that can take on only two possible values. These values are typically represented as 0 and 1, or as "success" and "failure," "yes" and "no," or any other meaningful pair of categories. Binary outcomes are often referred to as dichotomous variables.

Examples of binary outcome variables include:

Whether a customer will churn or not (0 for no churn, 1 for churn)

Whether a student will pass an exam or not (0 for fail, 1 for pass)

Whether a patient has a disease or not (0 for healthy, 1 for diseased).

1.11.9 Probability and Odds

In logistic regression, we are interested in estimating the probability of the binary outcome occurring given a set of predictor variables. The probability (p) is a value between 0 and 1, representing the likelihood of the outcome happening. It is often denoted as P(Y =1), where Y represents the binary outcome.

The odds of an event occurring are defined as the ratio of the probability of the event happening to the probability of it not happening. Mathematically, odds =p/(1 −p). Odds can range from 0 to infinity, where odds less than 1 indicate a lower probability, odds equal to 1 indicate an equal probability, and odds greater than 1 indicate a higher probability.

In logistic regression, we model the logarithm of the odds, known as the logit function, as a linear combination of the predictor variables. By transforming the probability into the logit, we ensure that the predicted values lie between negative and positive infinity, making them suitable for linear regression.

1.12 Which Regression to Use and When?

The choice of regression technique depends on various factors, including the nature of the data and the research objective. Here are ten cases and the corresponding regression techniques commonly used (Table 1.13):

1. Predicting House Prices: Multiple linear regression is often used when there are multiple predictors (e.g., square footage, number of bedrooms) to estimate the price of a house.
2. Forecasting Stock Market Returns: Autoregressive integrated moving average (ARIMA) models are suitable for time series data, making them valuable for predicting future stock market returns.

Table 1.13 Table to understand "When to use which method of regression?"

Types of regression	When to use?
Univariate	Only one quantitative response variable is present
Multivariate	Only two or more quantitative response variables are need
Simple	Only one predictor variable is needed
Multiple	Requires, two or more predictive variables
Linear	All parameters are linear. Sometimes we transfer the variable to make it linear
Non-linear	Relationship between response and predictor is linear
Analysis of Variance (ANOVA)	All predictors are quantitative variables

3. Analyzing Marketing Campaigns: Logistic regression is employed to model the relationship between binary outcomes (e.g., purchase vs. no purchase) and predictors (e.g., age, income) in marketing campaigns.
4. Predicting Student Performance: Support vector regression (SVR) is effective for predicting continuous outcomes, such as student test scores, based on various features like study hours and attendance.
5. Determining Credit Risk: Binary logistic regression is commonly used in credit risk assessment to predict the likelihood of loan default based on factors such as income, credit history, and loan amount.
6. Analyzing Customer Churn: Cox proportional hazards model is suitable for survival analysis and is often used to predict customer churn in industries like telecommunications or subscription-based services.
7. Estimating Sales Volume: Poisson Regression is useful when the dependent variable represents count data, such as the number of products sold, and predictors include factors like price and advertising expenditure.
8. Modeling Disease Progression: Generalized linear models (GLMs) are versatile and can handle a wide range of scenarios, including modeling disease progression based on factors like age, genetic markers, and lifestyle variables.
9. Forecasting Energy Consumption: Time series regression, such as the autoregressive integrated moving average with exogenous variables (ARIMAX) can be employed to predict energy consumption by incorporating external factors like weather conditions.
10. Predicting Customer Lifetime Value: Survival analysis, specifically the Kaplan–Meier estimator or Cox Proportional Hazards Model, can be used to estimate the expected lifetime value of customers based on their purchase history and time of churn.

Remember, the choice of regression technique may vary based on the specific dataset, assumptions, and objectives of the analysis. It is essential to consider the characteristics of your data and consult with domain experts to determine the most appropriate regression technique for your case.

1.13 Caution While Using Regression Analysis

While regression analysis is a powerful statistical tool for analyzing relationships between variables, it is important to exercise caution and be aware of potential pitfalls. Here are a few points to keep in mind when using regression analysis:

1. Causation versus correlation: Regression analysis can only establish correlations between variables and does not imply causation. Just because two variables are correlated does not mean that one causes the other. Additional evidence and careful interpretation are needed to establish causality.

2. Linearity assumption: Regression analysis assumes a linear relationship between the predictor variables and the response variable. If the relationship is nonlinear, the results may be misleading. Consider exploring alternative regression models or transforming variables to address nonlinearity.
3. Outliers: Outliers can have a significant impact on the regression model, affecting the slope, intercept, and overall fit. It's important to identify and handle outliers appropriately. Outliers may need to be removed or their effects mitigated through robust regression techniques.
4. Multicollinearity: When predictor variables are highly correlated with each other, multicollinearity occurs. This can lead to unstable coefficient estimates and difficulty in interpreting the results. Check for multicollinearity using diagnostic tests and consider methods such as variable selection or principal component analysis to address the issue.
5. Overfitting: Overfitting occurs when a regression model fits the noise or random fluctuations in the data rather than the underlying relationship. Overfit models may perform well on the training data but generalize poorly to new data. Regularization techniques like ridge regression or cross-validation can help mitigate overfitting.
6. Assumptions: Regression analysis relies on certain assumptions such as linearity, independence of errors, constant variance of errors, and normality of residuals. Violations of these assumptions can affect the validity of the results. Diagnostic tests and residual analysis should be performed to assess the model's assumptions.
7. Extrapolation: Regression models are reliable within the range of the observed data. However, extrapolating beyond that range is risky and may lead to unreliable predictions. Be cautious when making predictions outside the range of the data used to build the model.
8. Data quality: Regression analysis assumes that the data used is accurate, reliable, and representative. Carefully examine the data for missing values, measurement errors, and potential biases that could affect the analysis. Preprocess the data appropriately before conducting regression analysis.

It's important to approach regression analysis with a critical mindset and consider the limitations and assumptions involved. Additionally, consulting with experts or statisticians can help ensure accurate interpretation and appropriate use of regression analysis for your specific research or analysis.

Here are some real-life examples illustrating the cautionary points mentioned:

1. Causation versus correlation: A study finds a positive correlation between ice cream sales and crime rates. While the correlation is observed, it does not imply that ice cream consumption causes an increase in crime. The common underlying factor could be warmer temperatures, leading to both increased ice cream sales and higher crime rates.
2. Linearity assumption: A researcher assumes a linear relationship between advertising spending and sales. However, upon analyzing the data, they find that the relationship is curvilinear, with diminishing returns on advertising effectiveness

beyond a certain point. Failing to account for nonlinearity can lead to inaccurate predictions and decision-making.

3. Outliers: In a study analyzing the relationship between income and happiness, one participant reports an exceptionally high income. This outlier disproportionately influences the regression results, artificially inflating the estimated impact of income on happiness. Handling outliers appropriately is crucial to avoid biased or misleading conclusions.
4. Multicollinearity: In a regression model examining the factors influencing employee performance, it is discovered that job satisfaction and work-life balance are highly correlated. This multicollinearity can make it challenging to isolate the individual effects of each variable accurately, potentially leading to unreliable coefficient estimates.
5. Overfitting: A data scientist develops a regression model to predict stock prices based on various economic indicators. By including a large number of predictors, the model achieves near-perfect fit on historical data. However, when tested on new data, the model fails to generalize well, highlighting the problem of overfitting.
6. Assumptions: In a study investigating the impact of class size on academic performance, the residuals of the regression model exhibit heteroscedasticity, meaning the variability of errors increases as class sizes become larger. This violates the assumption of constant variance, and the model's estimates and statistical inferences may become unreliable.
7. Extrapolation: A company uses a regression model to forecast product demand based on historical sales data. However, when the model is applied to predict sales in a new market where consumer behavior differs significantly, the forecasts prove inaccurate due to extrapolating beyond the range of observed data.
8. Data quality: A researcher conducts a regression analysis on a dataset containing missing values. Failing to appropriately handle these missing values can introduce bias and affect the accuracy of the results. Imputing missing data using appropriate methods or removing observations with missing values can help ensure the integrity of the analysis.

These examples illustrate the importance of understanding the limitations and potential pitfalls associated with regression analysis, and the need for careful consideration and scrutiny in real-life applications.

1.14 Outliers in Regression Analysis

Regression analysis is a powerful statistical tool used to establish relationships between variables and predict outcomes. However, in real-world data, outliers can significantly impact the accuracy and reliability of regression models. Outliers are data points that deviate significantly from the general trend of the dataset and can distort the estimated relationships between variables. Outliers can significantly affect

the validity and reliability of regression analysis. Understanding the causes and consequences of outliers is crucial for researchers and data analysts to make informed decisions regarding outlier handling. Employing appropriate outlier detection techniques and implementing suitable remedies can lead to more accurate regression models and more reliable predictions, enhancing the overall quality of statistical analyses. Mathematically, we can say that the response variable "y" does not follow the general trend of rest of data. Outliers will lead to error in the regression equation as a result of which these lines will not be accurate in predicting other data values.

1.14.1 Causes of Outliers in Regression Analysis

Outliers can arise due to various reasons, including data entry errors, measurement inaccuracies, natural variability, or extreme events. They can be caused by random chance or may indicate genuine deviations in the data. Outliers can be univariate (occurring in a single variable) or multivariate (occurring in multiple variables simultaneously).

1.14.2 Impact of Outliers on Regression Analysis

Outliers can have several adverse effects on regression analysis:

(a) Biased Estimates: Outliers can substantially influence the slope and intercept of the regression line, leading to biased coefficient estimates.
(b) Reduced Model Fit: Outliers can introduce heteroscedasticity (unequal variance) and nonlinearity, affecting the assumptions of regression models and reducing the overall goodness-of-fit.
(c) Inaccurate Predictions: The presence of outliers can lead to inaccurate predictions, as the regression model may prioritize fitting these extreme values rather than the underlying pattern of the data.
(d) Decreased Statistical Significance: Outliers can inflate the standard errors of coefficient estimates, reducing the statistical significance of the relationships between variables.

1.14.3 Detecting Outliers in Regression Analysis

Several methods can be used to identify outliers in regression analysis:

(a) Residual Analysis: Plotting the residuals (the differences between observed and predicted values) can reveal patterns in the data, including outliers.

(b) Cook's Distance: Cook's distance measures the influence of each observation on the entire regression model. Large Cook's distances indicate potential outliers.
(c) Z-Scores: Calculating the Z-scores for each data point can help identify extreme values that deviate significantly from the mean.
(d) Mahalanobis Distance: This metric measures the distance of each data point from the centroid of the data, considering correlations between variables.

1.14.4 Handling Outliers in Regression Analysis

Once outliers are detected, various strategies can be employed to mitigate their impact on regression analysis:

(a) Transformation: Applying transformations to the data, such as logarithmic or power transformations, can reduce the impact of outliers and improve model fit.
(b) Robust Regression: Robust regression techniques, like the Huber or Tukey bisquare methods, give less weight to outliers during the estimation process.
(c) Data Cleaning: If outliers are due to data entry errors, correcting or removing these data points may improve the accuracy of the analysis.
(d) Data Stratification: Separating the data into subgroups based on certain characteristics can help handle outliers in specific subsets without affecting the entire dataset.

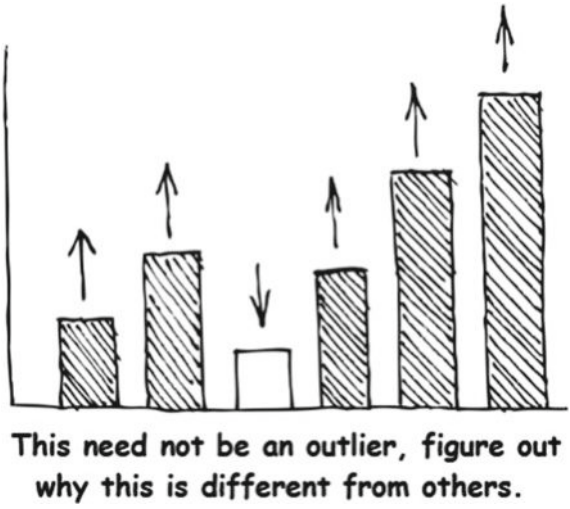
1.14.5 Removing Outliers on Regression Lines

Removing outliers is a common practice in data analysis and regression modeling to improve the accuracy and reliability of the regression line. Outliers are data points that significantly deviate from the overall pattern of the data and can have a substantial impact on the regression line's slope and intercept.

Here's a step-by-step guide on how to remove outliers from a regression analysis:

1. Visualize the Data: Plot the data points on a scatter plot to identify any potential outliers. Outliers are usually data points that are far away from the general trend of the data.
2. Set Criteria for Identifying Outliers: There are several methods to identify outliers, such as using the Z-score or the interquartile range (IQR). The Z-score measures how many standard deviations a data point is from the mean, while the IQR is the range between the 25th and 75th percentiles. Data points that fall outside a certain range of Z-scores or outside the IQR can be considered outliers.
3. Calculate the Regression Line: Before removing outliers, calculate the initial regression line using all data points. This will serve as a baseline for comparison after removing the outliers.
4. Identify Outliers: Apply the chosen outlier detection method to the data and identify the data points that meet the criteria for being outliers.

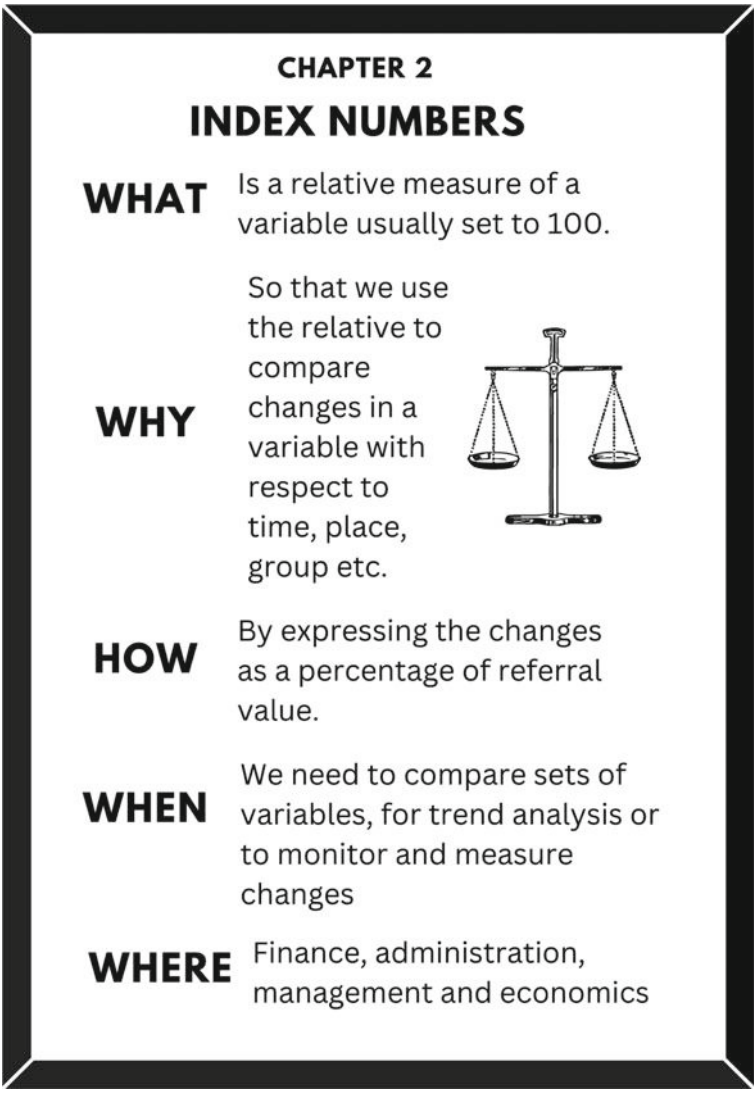
Fig. 1.4 Outliers



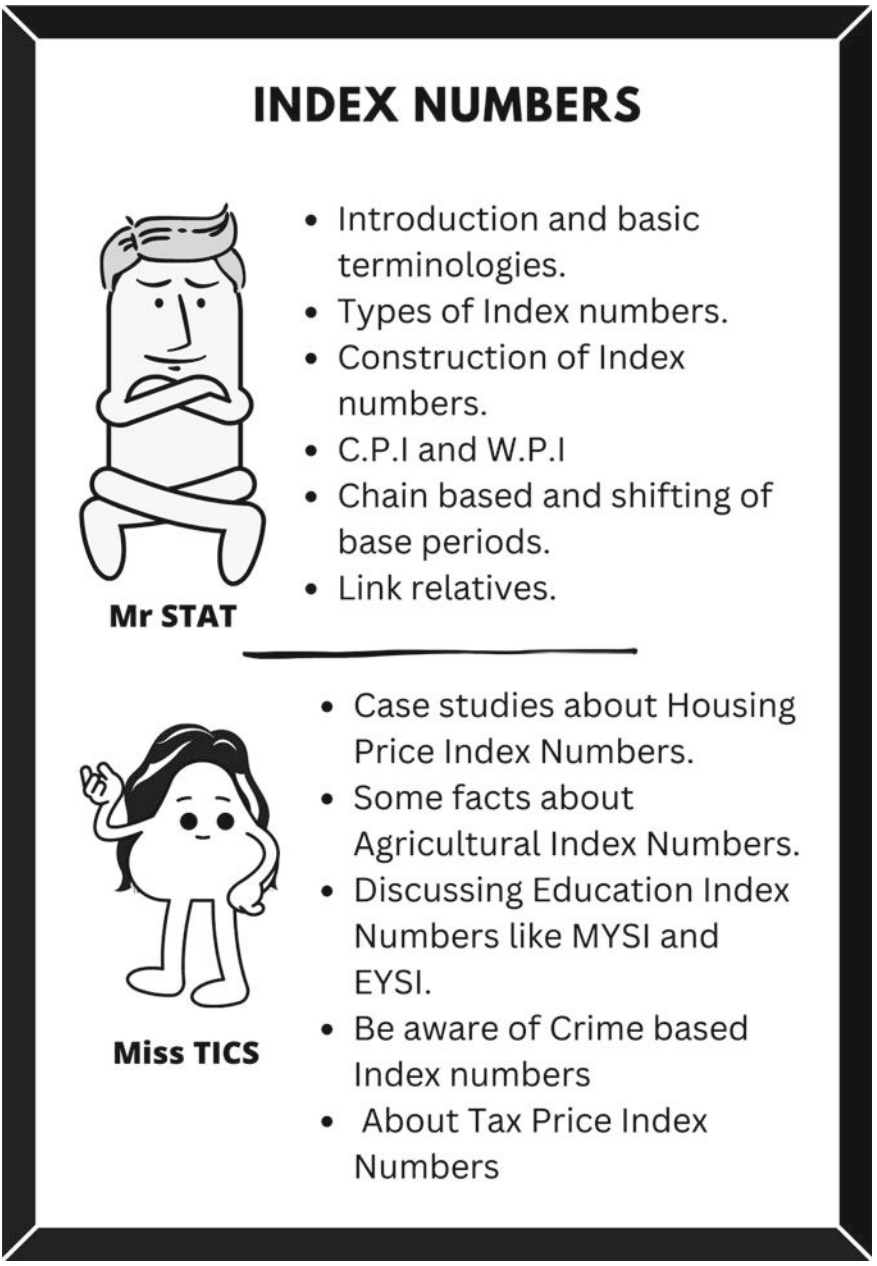
5. Remove Outliers: Once you have identified the outliers, remove them from the dataset. Depending on the context and the data, you can either exclude them completely or replace them with more appropriate values (e.g., mean, median).
6. Recalculate the Regression Line: With the outliers removed, calculate the regression line again using the modified dataset.
7. Assess Model Performance: Compare the performance of the regression model before and after removing the outliers. Common metrics like R-squared, mean squared error (MSE), or root mean squared error (RMSE) can be used for evaluation. If the model's performance improves significantly after removing the outliers, it indicates that the outliers were indeed affecting the model's accuracy.
8. Interpret Results: Once you have obtained a more reliable regression line, interpret the results and draw conclusions based on the updated model (Fig. 1.4).

It's important to note that the decision to remove outliers should be made judiciously. Sometimes outliers represent genuine data points, extreme values, or unique events that are essential for understanding the underlying phenomenon. Removing outliers indiscriminately without a proper justification can lead to biased and inaccurate results. Always exercise caution and domain knowledge when handling outliers in data analysis.

Chapter 2
Index Numbers



S. Prasad, *Advanced Statistical Methods*, https://doi.org/10.1007/978-981-99-7257-9_2



What are Index Numbers?

Index numbers are a type of economic indicator that represent the relative change in the value of a variable (e.g., price, quantity, production, income, etc.) compared to a

base period or a reference point. They are used to simplify complex data and present it in a more understandable and comparable form.

Why Use Index Numbers?

- Comparative analysis: Index numbers allow you to compare different time periods or geographical regions in a standardized manner. This helps in understanding the magnitude and direction of changes.
- Tracking changes over time: Index numbers help track trends and identify patterns, such as inflation, economic growth, market performance, and consumer behavior.
- Forecasting: By analyzing historical index data, you can make predictions about future trends and potential outcomes.
- Policy decisions: Governments and policymakers use index numbers to assess the effectiveness of their policies and make data-driven decisions.
- Business performance: Companies use index numbers to evaluate their financial performance, monitor market trends, and benchmark against competitors.

How to Calculate Index Numbers?

To calculate an index number, follow these steps:

- Select a base period and assign it a value of 100. This period acts as a reference point.
- Collect data for the variable of interest for the current period and the base period.
- Divide the current value by the base period value and multiply by 100 to get the index number.

Mathematically, the formula is: Index number =(Current value/Base period value) $\times$100.

When to Use Index Numbers?

You can use index numbers in various scenarios:

- Tracking inflation: Consumer Price Index (CPI) and Producer Price Index (PPI) are used to monitor changes in price levels over time.
- Economic indicators: Index numbers like the Gross Domestic Product (GDP) and Industrial Production Index help gauge economic performance.
- Financial market analysis: Index numbers, such as stock market indices (e.g., S&P 500, Dow Jones), show the overall performance of the market.
- Performance evaluation: Companies use indices to assess sales, production, and other key performance indicators.

In summary, index numbers are a valuable tool for comparing data, analyzing trends, making predictions, and aiding decision-making across various fields. They are especially useful when dealing with complex datasets and time-series analysis. However, it is essential to understand their limitations and context-specific applicability before interpreting and using them to draw conclusions.

2.1 Introduction

Index numbers are statistical measures used to track changes over time in a particular variable or group of variables.

Few characteristics are:

1. Index numbers are specialized averages as they can be used to compare different types of items with different units. They are calculated by units of consumption, rather than units of measurement. For example, milk, oil, etc., are measured in liters, rice, wheat, etc., in terms of kgs, and eggs, etc., per dozen. We can use index numbers though all these are in different units.
2. Index numbers are expressed in percentages of relative changes, but the sign "%" is not used.
3. Index numbers measure those changes which cannot directly be measured, e.g., "price level," "Economic activity," "Cost of living," etc.
4. Index numbers are meant for comparing over made over different intervals of time with reference to a particular base year.
5. Index numbers have universal application as they are used to ascertain changes in different sectors of study.
6. Index numbers are also an example of summary statistics.
7. They are simple tools measuring relative changes and are expressed in terms of percentage.
8. The index number is usually expressed as 100 times the ratio to the base value.
9. Comparisons can also be made with multiple entities with different unit values, thus making index numbers a specialized average.

2.2 Definitions

Index numbers measure the net change in the magnitude of a set of related variables. Some popular definitions are as follows:

Dr. A. L. Bowely: Index number is a statistical tool that measures changes in a variable or a group of variables over time or other aspects.

Wessel, Willett, *and Simone*: An index number is a special type of average that provides a measurement of relative changes from time to time or from place to place.

Croxton and Cowden: Index numbers are devices for measuring differences in the magnitude of a group of related variables.

Spiegel: An index number is a statistical measure designed to show changes in a variable or a group of related variables with respect to time, geographic location, or other characteristics.

2.3 Important Uses of Index Numbers

1. Index numbers measure changes in price level and indicate inflationary or deflationary tendencies in the data.
2. Index numbers help in calculating the real value of money or purchasing power of money.
3. Index numbers can be used to make measuring adjustments in the wages of employees. The wages of employees can be increased as per the increase in the Cost-of-Living index number.
4. Index numbers guide economic and business policies. For example, the relative changes in production are given by index numbers of industrial production.
5. Trends and tendencies of various phenomena over time are given by index. For example, index numbers can be used to track changes in various economic indicators, such as inflation, GDP, stock market performance, and consumer spending. They can also be used to analyze changes in other fields, such as education, health, and demographics.
6. By using index numbers, researchers and analysts can identify patterns and trends in the data that might not be immediately apparent from looking at the raw data. This information can be used to make informed decisions about future policies or investments and to understand how various phenomena are changing over time.

2.3.1 Index Numbers in Analytics

In analytics, comparing data based on time, quality, quantity, money, and value is a very common practice. For example, an agronomist decides to analyze soil degradation on a piece of land. Certain aspects of his study may include,

– A check on the alkalinity, acidity, and salinity of soil with respect to a specific period.
– The quality of production of the main crop and the alternative crops that are grown across the year.
– The number of fertilizers and chemicals used on the soil to decipher the rate of growth of microorganisms that destroy the crop.
– Change in the number of crimes after a new law has been introduced, etc.

We use index numbers to calculate the rate of change in various fields in different variables, over time, place, and categories. Index numbers provide a way to compare the relative changes in a variable over different periods or locations. They are often used to track changes in the prices of goods and services, but can also be used to track changes in other variables such as production levels, employment rates, and economic growth.

2.3.2 Index Numbers in Nation Building

Index numbers play a crucial role in nation building by providing valuable information and insights for policymakers, businesses, and citizens. Here are some ways in which index numbers contribute to nation building:

1. Economic Monitoring and Policy Formulation: Index numbers, such as the Consumer Price Index (CPI), Wholesale Price Index (WPI), and Gross Domestic Product (GDP) growth rate, help monitor and assess the overall economic health of a nation. Policymakers rely on these indicators to understand trends, identify areas of concern, and formulate appropriate economic policies. Index numbers provide insights into inflation, economic growth, productivity, and other key economic indicators, helping policymakers make informed decisions to foster sustainable economic development.
2. Assessing Standard of Living: Index numbers related to household income, poverty rates, and living standards provide valuable insights into the well-being of citizens. These indicators help policymakers understand the distribution of income, identify vulnerable populations, and design targeted policies to alleviate poverty, improve living conditions, and promote social development. By monitoring and analyzing index numbers related to standard of living, nations can strive for inclusive growth and ensure the welfare of their citizens.
3. Sectoral Analysis and Resource Allocation: Index numbers enable policymakers to analyze and understand the performance of different sectors within the economy. For instance, the Industrial Production Index (IPI) provides insights into the manufacturing sector's performance, while the Agricultural Price Index (API) tracks price changes in the agricultural sector. By examining these sectoral index numbers, policymakers can allocate resources, implement targeted policies, and promote balanced development across various sectors, ensuring sustainable economic growth and job creation.
4. Business Decision-Making: Index numbers provide crucial information for businesses to make informed decisions. For instance, the Business Confidence Index (BCI) helps gauge the sentiments and expectations of businesses, enabling them to plan investments, expand operations, or adjust strategies based on the prevailing economic conditions. Additionally, index numbers related to consumer spending, purchasing power, and market demand assist businesses in developing pricing strategies, forecasting demand, and identifying market opportunities.
5. Investor Confidence and Economic Development: Reliable index numbers foster investor confidence and attract foreign direct investment (FDI). Investors use these indicators to assess the economic potential and stability of a nation. Index numbers such as GDP growth rate, inflation rate, and ease of doing business rankings influence investment decisions. By maintaining accurate and transparent index numbers, nations can attract investment, promote economic development, and create employment opportunities.
6. Monitoring Sustainable Development Goals (SDGs): Index numbers are instrumental in tracking progress toward the Sustainable Development Goals (SDGs)

outlined by the United Nations. These goals encompass areas such as poverty eradication, education, healthcare, gender equality, and environmental sustainability. Index numbers related to these areas provide a quantitative measure of progress, enabling policymakers to identify gaps, prioritize interventions, and track the impact of policies and initiatives aimed at achieving the SDGs.

Thus, index number is a reliable tool used for quick and easy comparisons. This tool is predominantly used in all fields of study due to its practicality, importance, and simplicity.

2.3.3 Index Numbers Are Economic Barometers

A lot of econometric concepts use this tool to compare, analyze, and understand trends of many economic factors of life and industries.

Index numbers are often used as economic barometers because they provide a useful way to track changes in economic variables over time. An index number is a statistical measure that compares a current value of a variable to its value at some base period, which is typically set to 100. By comparing current values to a base period, index numbers allow us to see how the variable has changed over time relative to its original level. Some examples of economic variables that are often tracked using index numbers include:

1. Inflation: Inflation is the rate at which the general level of prices for goods and services is rising. The Consumer Price Index (CPI) is one commonly used index number that tracks changes in the cost of living over time.
2. Stock market performance: The Dow Jones Industrial Average and the S&P 500 are two index numbers that are commonly used to track changes in the stock market over time.
3. Gross Domestic Product (GDP): GDP is the total value of goods and services produced within a country's borders in a given period. GDP index numbers can be used to track changes in economic growth over time.

By tracking changes in these and other economic variables using index numbers, policymakers, investors, and others can gain insights into the overall health of the economy and make informed decisions about future actions.

Some case studies on index numbers as economic barometers:

1. Purchasing Managers' Index (PMI) as an Economic Barometer: The PMI is an index that provides insight into the manufacturing sector's economic activity. It measures factors such as new orders, production levels, employment, supplier deliveries, and inventories. During the global financial crisis of 2008–2009, the PMI numbers dropped significantly worldwide, indicating a contraction in manufacturing activity. This decline in the PMI served as a barometer for the overall health of the economy, reflecting the impact of the crisis on businesses and consumer demand.

2. Leading Economic Index (LEI) as an Economic Barometer: The LEI is a composite index that combines multiple economic indicators to provide an overall picture of the economy's direction. It aims to predict changes in economic activity before they occur. A case study can focus on the United States and the Great Recession of 2007–2009. The LEI declined before the recession officially began, serving as a barometer for the impending economic downturn. By analyzing the leading economic indicators, policymakers, economists, and businesses can anticipate shifts in economic activity and take proactive measures.
3. Stock Market Indices as Economic Barometers: Stock market indices, such as the S&P 500 or the Dow Jones Industrial Average, are often considered barometers of economic health. They provide insights into investor sentiment and overall market conditions. A case study could examine the relationship between stock market indices and economic recessions. For example, during the dot-com bubble burst in the early 2000s, stock market indices experienced significant declines, signaling an economic slowdown. These index movements acted as barometers, reflecting the market's assessment of the economy.
4. Consumer Confidence Index (CCI) as an Economic Barometer: The CCI measures consumers' sentiments and expectations regarding the overall state of the economy. It is an essential index for understanding consumer behavior and economic activity. A case study could focus on the impact of the COVID-19 pandemic on consumer confidence and economic barometers. As the pandemic unfolded, consumer confidence declined significantly in many countries, indicating a decrease in consumer spending and overall economic activity. The CCI served as a barometer for the economic impact of the pandemic on consumer behavior and sentiment.

These case studies demonstrate how various index numbers act as economic barometers, providing valuable insights into economic activity, market conditions, and consumer sentiment. By monitoring and analyzing these indices, policymakers, economists, and businesses can make informed decisions, anticipate economic shifts, and take timely measures to mitigate risks or leverage opportunities.

2.3.4 Index Numbers and Agriculture

The Ministry of Statistics and Program states that index numbers are very useful in the prediction of the agricultural progress of the country which is one of the vital and primary sources of income for the nation. This department uses index numbers to study the trends over time with respect to:

- Type of the crop.
- The area under cultivation.
- The quantity of yield.

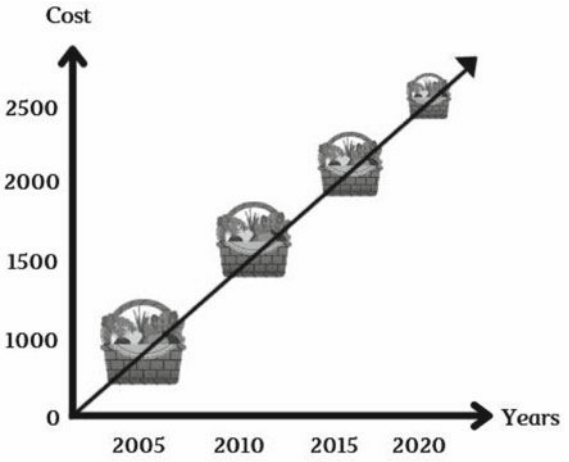


Fig. 2.1 Depicting purchasing power of money

- The productivity of the crop.
- Soil fertility.
- Prices/profits earned.

Besides this, the department also calculates index numbers in terms of trade between agricultural and non-agricultural sectors for a better understanding of the market.

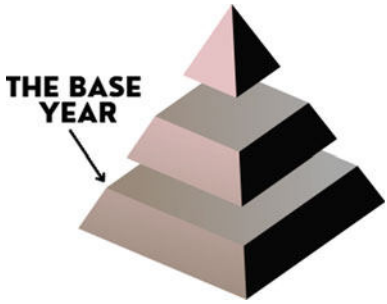
Value index numbers are one of the frequently used index numbers to understand the value (quantity * price) of money which is also called the purchasing capacity with respect to time. Figure 2.1 aims to depict the purchasing abilities of groceries over a period of time. It is an example of how price and quantity are inversely proportional to each other.

2.4 The Base Year

The values of the year which are considered as a base for comparison are called the base year. For example, India conducts a decennial (once in 10 years) census. The 16th Indian census was in the year 2021. Therefore, comparisons can be made based on basic population characteristics such as household composition, expenditure, income, and size. These values are compared with the previous census values of the year 2011. So, in this case, 2011 becomes the base year and 2021 becomes the current year values (Fig. 2.2).

- The base and current years can be chosen by the analyst at his convenience.
- The base year value is set to 100 by default.

Fig. 2.2 Base year



- Base year values are represented by 0 and current year values are represented by 1.
- Example: If P denotes price and Q denotes quantity, P_1 refers to the price of the current year and Q_o refers to the quantity value of the base year.
- The notation P_{01} refers to the price index number of the current year.

Example 1 These are the prices of a microscope of a lab. Calculate the index numbers considering (Table 2.1).

a. 1981 as base period.
b. Base period as 1974–1976.

Table 2.1 Prices of microscope from 1974 to 1982

Year	1974	1975	1976	1977	1978	1979	1980	1981	1982
Price of commodity	1200	1500	1235	1345	1455	1290	1546	1234	1654

Table 2.2 Index numbers with base period as 1981 and 1974–1976 as a base

Year	Price of commodity	Index number 1981 = 100	Index numbers with 1974–1976 as base
1974	1200	$\frac{1200}{1234} * 100 = 97.24$	$\frac{1200}{3935} * 100 = 30.49$
1975	1500	$\frac{1500}{1234} * 100 = 121.55$	$\frac{1500}{3935} * 100 = 38.11$
1976	1235	$\frac{1235}{1234} * 100 = 100.08$	$\frac{1235}{3935} * 100 = 31.38$
1977	1345	$\frac{1345}{1234} * 100 = 108.99$	$\frac{1345}{3935} * 100 = 34.18$
1978	1455	$\frac{1455}{1234} * 100 = 117.90$	$\frac{1455}{3935} * 100 = 36.97$
1979	1290	$\frac{1290}{1234} * 100 = 104.53$	$\frac{1290}{3935} * 100 = 32.78$
1980	1546	$\frac{1546}{1234} * 100 = 125.28$	$\frac{1546}{3935} * 100 = 39.28$
1981	1234	$\frac{1234}{1234} * 100 = 100$	$\frac{1234}{3935} * 100 = 31.35$
1982	1654	$\frac{1654}{1234} * 100 = 134.03$	$\frac{1654}{3935} * 100 = 42.03$

Table 2.3 Performance index of various sectors of economy

Year	Agriculture	Manufacture	Construction	Retail	Hospitality
2010	100	100	100	100	100
2011	102.6	104	105.3	103	100.1
2012	109.9	110.7	118.1	111.8	100.9
2013	118.4	106.5	115.2	120.3	101.5
2014	121.8	105.1	124.1	121.7	102.1
2015	125.6	101.2	126.3	125.4	102

Solution

The general formula is given as, Index Number $= \frac{\text{Price of current year}}{\text{Price of base year}} * 100$.

Price of the base period 1974–1976 will be $1200 + 1500 + 1235 = \mathbf{3935}$.

Example: Consider the following example where the mayor at a conference presents the earnings in different sectors of the economy in a city. The year 2010 is the base year with the index of all sectors being 100.

The mayor concludes that the earnings from the agriculture and retail industry are progressive, but the manufacturing and hospitality industry earnings are poor and are declining. Let's calculate the percentage changes to check. For the year 2015, the income from the agriculture industry has progressed by 22.41% $\left(\frac{125.6-102.6}{102.6}\right)$, whereas the hospitality sector has scaled up by just 1.89% $\left(\frac{102-100.1}{100.1}\right)$ over a period of 5 years starting from 2011. This was simple to calculate and decipher only because the values in the tables are the index values. Table 2.2 depicts 102.6 and 104 for the year 2011 under the agriculture and manufacturing departments, respectively. This is a comprehensive value of all the factors of assessment such as social, economic, and financial aspects of the industry. Therefore, an index is a composite statistic. Therefore, it is an aggregate of many indicators (Table 2.3).

2.5 Types of Index Numbers Based on Methods of Calculation

1. **Relative index numbers**: In this method, the price of each item in the current year is expressed as a percentage of the price in the base year.

$$\text{Price Index}_{\textbf{Relative method}} = \frac{\textbf{Price in the given year}}{\textbf{Price in the base year}} * \mathbf{100} = \frac{\boldsymbol{P}_1}{\boldsymbol{P}_0} * \mathbf{100}$$

The best example of a simple relative average is the Wholesale Price Index or W.P.I.

Suppose the WPI is 128%, this simply means that the value of a representative basket of wholesale goods has increased by 28% with respect to the price of the base year.

2. **Aggregate Index numbers**: As the name suggests, this is a comprehensive value of all the sub-entities. An aggregate price index number refers to a single value that includes prices of all sub-items in a given year and it is expressed in terms of the percentage of the base year. In this method, equal importance is given to all the sub-items and the prices are quoted in the same units.

$$\boldsymbol{Price\ Index}_{\boldsymbol{Aggregate\ method}} = \frac{\textbf{Aggregate price in the given year}}{\textbf{Aggregate price in the base year}} * \mathbf{100} = \frac{\sum \boldsymbol{P}_1}{\sum \boldsymbol{P}_0} * \mathbf{100}$$

where $\sum P_1$ and $\sum P_0$ are the total prices of various commodities in the current and base year, respectively.

3. **Weighted index numbers**: A type of index number in which rational weights are assigned to various entities as per their importance. It is calculated as, the ratio between the summation of the product of weights with price relatives and the summation of the weights.

$$\textbf{Weighted price index numbers} = \frac{\Sigma \mathbf{WP}}{\Sigma \mathbf{W}} * \mathbf{100}$$

Industrial Production Index (I.P.I) is the best example of the weighted average of quantity relatives. I.I.P reveals the change in the industrial production and is defined as,

I.I.P$_{01} = \frac{\sum q_1 * W}{\sum W}$, where q_1 is the quantity of the current year and W is the weight of the entity.

Example: Consider the following simple example to get a better understanding of the relative and aggregative method of calculating index numbers (Tables 2.4 and 2.5).

The simple average price relative of commodities is calculated as:

$$P_{01} = 746.63/4 = \mathbf{186.65}$$

Table 2.4 Relative and aggregate method of calculating index numbers

Commodities	Base price	Current price	Calculation
A	9	18.5	(18.5/9) * 100 =205.55
B	14	27.8	(27.8/14) * 100 =198.57
C	37	62.5	(62.5/37) * 100 =168.91
D	26.5	46	(46/16.5) * 100 =173.58
Total	86.5	154.8	746.63

The aggregate price relative of commodities is calculated as:

$$P_{01} = \frac{\sum P_1}{\sum P_0} * 100 = (154.8/86.5) * 100 = \mathbf{178.95}$$

Example: A local newspaper about prices of limestone for construction purpose stated that, "In 1996 the average price of a commodity was 25% more than in 1995, but 20%less than in 1994 and it was 70% more than in 1997."

a. Let us simplify the information and tabulate it for better understanding.
b. Also let us rewrite the information with price relatives using 1995 as base period.
c. How are the index numbers changing if we consider the average price of 1994 and 1995 as base?

Solution

a. Let us consider the price of limestone for the year 1996 is 100.

Therefore, for year 1995, the prices are expected to rise by 25% which means,

$$\text{Price in } 1995 = \frac{100}{125} \times 100 = 80.$$

For the year 1994, is 20% less, which means,

$$\text{Price in } 1994 = \frac{100}{80} \times 100 = 125$$

For the year 1997, is 70% more, which means,

$$\text{Price in } 1997 = \frac{100}{170} \times 100 = 58.82.$$

b. Price relatives using 1995 as base period will be,

$$P_{01}(1994) = \frac{125}{80} \times 100 = 156.25$$

Table 2.5 Price index of limestone

Year	Price index	Index numbers Base (1995)	Index numbers Base 102.5
1994	125	156.25	121.95
1995	80	100	78.05
1996	100	125	97.56
1997	58.82	73.52	57.39

$$P_{01}(1995)=100$$

$$P_{01}(1996)= \frac{100}{80} \times 100 = 125$$

$$P_{01}(1997) = \frac{58.82}{80} \times 100 = 73.52$$

c. Average price of 1994 and 1995 $= \frac{80+125}{2} = 102.5$

If 102.5 is taken as a base average, then the price relatives would be:

$$P_{01}(1994)= \frac{125}{102.5} \times 100 = 121.95$$

$$P_{01}(1995) = \frac{80}{102.5} \times 100 = 78.05$$

$$P_{01}(1996)= \frac{100}{102.5} \times 100 = 97.56$$

$$P_{01}(1997)= \frac{58.82}{102.5} \times 100 = 57.39$$

Consider the tabulated information, for a better understanding.

2.6 Price Relative

A price relative is simply a ratio between current and base year prices expressed in terms of percentages. These price relatives are then averaged to get the index number. The average chosen could be arithmetic mean, geometric mean, or even median.

The most commonly used weights are the product of the value of the base year which is p_0q_0. The weighted average of price relative is when weights are attached to the unweighted price relatives. The weighted average using simple arithmetic mean is given as:

$$P_{01} = \frac{\sum \left[\frac{p_1}{p_0} * 100 \right] * p_0 q_0}{\sum p_0 q_0} = \frac{\sum wP}{\sum w}$$

Here, price relative $P = [p_1/p_0] * 100$ and weights $w = p_0 q_0$.

The weighted average using geometric mean is given as:

$$P_{01} = \text{antilog}\left(\frac{\sum w \log P}{\sum w}\right).$$

2.6.1 The Price, Quantity, and Value Index Numbers

Example: A trader made note of the price and quantity of 10 ml of unadulterated eucalyptus oil that he sold for 5 years. After two years of selling, he felt that the product will continue to have more demand even if the price was increased to 65 from 52. Do you think that the price rise was justified? Compute the price, quantity, and value index and also interpret the results (Table 2.6).

Solution

By default, the initial year of study is considered to be the base year, and thus, the price and quantity values of 2015 will be considered as the base values for calculations. The value of an entity is defined as the product of the price and quantity of the entity.

Price Index for 2016 =(52/50) *100, for 2017 is (65/50) *100, for 2018 =(65/50) *100, and so on. Similarly, quantity index for 2016 =(240/200) *100, for 2017 =(120/200) *100, and so on. Thus, the value index is also computed in the same manner.

Interpretation: Only the quantity index for the years 2017 and 2018 is less than 100. This means that the quantity sold is 40% and 5% lesser than the base period of 2015. One of the probable reasons could be the sudden increase in the price, yet over a period of time, the value of the product was understood, and thus, the quantity index increased as the demand rose. Hence, the price rise was justified. The value index however remained very positive and progressive throughout the period of study.

Table 2.6 Price and quantity index calculation for eucalyptus oil

Year	Price	Price index	Quantity	Quantity index	Value	Value index
2015	50	100	200	100	10,000.00	100
2016	52	↑104	240	↑120	12,480.00	↑124.8
2017	65	↑130	120	↓60	12,320.00	↑123.2
2018	66	↑132	190	↓95	18,000.00	↑180
2019	75	↑150	235	↑117.5	20,480.00	↑204.8

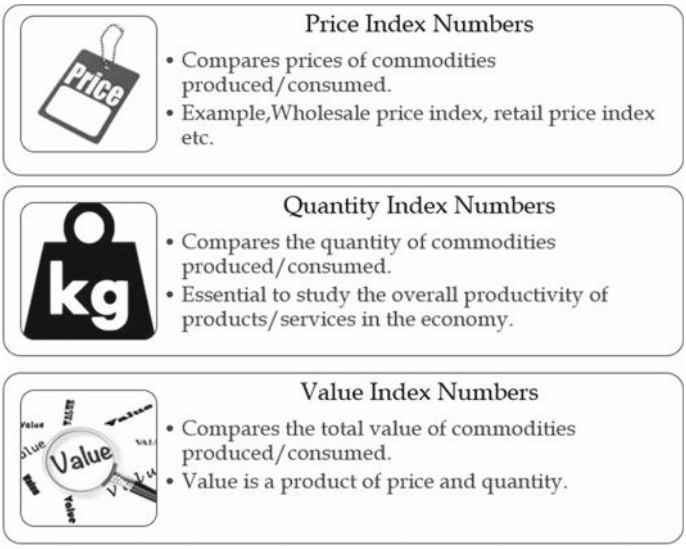


Fig. 2.3 Price, quantity, and value index numbers

2.7 Consumer Price Index Number—C.P.I.

CPI is a very extensively used economic indicator of price change. Economists have consolidated a set of goods that are often purchased and termed it "the basket of goods." CPI refers to the rate of change in the prices of a basket of goods. Thus, CPI is also called the Cost-of-Living index as it is used to understand the consumption pattern and composition of all household items. The unweighted CPI formula is (Figs. 2.3 and 2.4)

$$\text{C.P.I.} = \frac{\text{Cost of a basket of goods in a given year}}{\text{Cost of a basket of goods in the base year}} * 100$$

In the basket of goods, the commodities are categorized into different groups and weights are attached based on their importance.

Therefore, weighted price index numbers $= \frac{\sum WP}{\sum W} * 100$

where $P = [p1/p0]^* 100$ and W is the weights.

For example, if we say, CPI with the base year of 2000 is 560 in August 2011. We mean that a person buying a certain essential basket of commodities paid 100 rupees in the year 2000 now pays 560 rupees for the same basket of commodities in August 2011. A general increase in prices in an economy is called inflation. Let's understand CPI and inflation with a simple example. Consider the situation where we pack picnic bags for kids in the year 1995 and 2005 (Table 2.7).

Total cost (1995) would be:

Fig. 2.4 Comic on inflation by Walt Handlesman

Table 2.7 Price and quantity of junk foods such as chips, cooldrinks, and chocolates

	Chips	Cooldrinks	Chocolates
Price 1995	5	15	10
Quantity 1995	3	2	5

$$(5 * 3) + (15 * 2) + (10 * 5) = 15 + 30 + 50 = 95 \text{ rupees.}$$

Now after 10 years, we still pack the bag with the same items and quantity, but the prices would have certainly increased. Therefore, chips, cool drinks, and chocolates now cost 10, 25, and 30 rupees, respectively.

Total cost (2005) would be:

$$(10 * 3) + (25 * 2) + (30 * 5) = 30 + 50 + 150 = 230 \text{ rupees.}$$

C.P.I. is just an index value, indexed to 100 in the base year. To compute C.P.I, let us consider the year 1995 as the base year and 2005 will be the current year.

$$\text{C.P.I.}(2005) = (230/95) * 100 = 242.10.$$

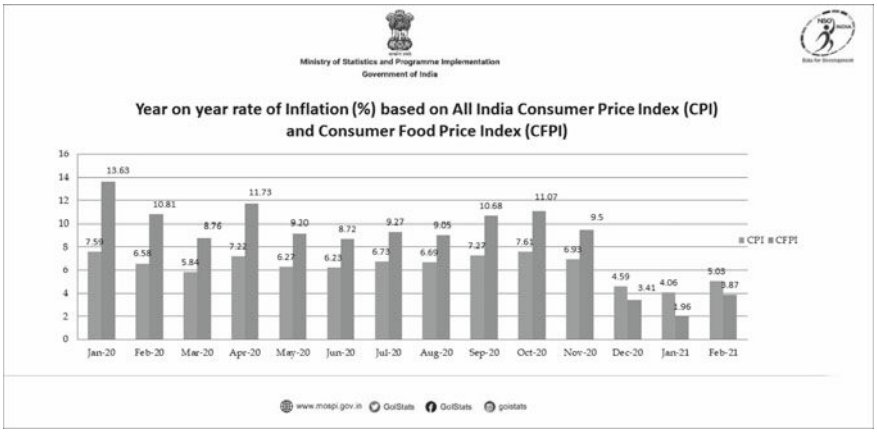


Fig. 2.5 Which shows the C.P.I. and Consumer Food Price Index (C.F.P.I) values on a monthly basis

2.7.1 How is the CPI Market Basket Determined?

The CPI market basket is developed from detailed expenditure information provided by families and individuals on what they actually bought. There is a time lag between the expenditure survey and its use in the CPI. For example, CPI data in 2020 and 2021 was based on data collected from the Consumer Expenditure Surveys for 2017 and 2018. In each of those years, about 24,000 consumers from around the country provided information each quarter on their spending habits in the interview survey. To collect information on frequently purchased items, such as food and personal care products, another 12,000 consumers in each of these years kept diaries listing everything they bought for 2 weeks. Over the 2 years, the expenditure information came from approximately 24,000 weekly diaries and 48,000 quarterly interviews used to determine the importance, or weight, of the item categories in the CPI index structure. (https://www.bls.gov/cpi/questions-and-answers.htm) (Fig. 2.5).

This is the data from the Ministry of Statistics and Programme Implementation., Government of India. Which shows the C.P.I. and C.F.P.I. (Consumer Food Price Index) values on a monthly basis. https://www.rba.gov.au/education/resources/digital-interactives/inflation-explorer/.

2.7.2 Some Case Studies on Consumer Price Index (CPI) Numbers

1. The United States CPI during an Economic Downturn: During the global financial crisis of 2008–2009, the United States experienced a significant economic downturn. The CPI reflected this downturn, as consumer prices decreased due to

decreased demand and deflationary pressures. The CPI numbers showed a decline in prices for various goods and services, including housing, transportation, and consumer durables. This information was crucial for policymakers to understand the extent of the economic crisis and make informed decisions to stimulate the economy.

2. Inflation in Venezuela: Venezuela has faced a severe economic crisis marked by hyperinflation in recent years. The country's CPI numbers skyrocketed as a result, reflecting the rapid increase in prices. In 2018, the inflation rate in Venezuela reached an astronomical level, with the monthly CPI increasing by hundreds or even thousands of percentage points. These alarming CPI numbers highlighted the economic instability and the challenges faced by the Venezuelan population in affording basic necessities.
3. CPI and Price Changes in Japan: Japan experienced a prolonged period of deflation, known as the "Lost Decade" in the 1990s and early 2000s. The CPI numbers during this period showed a consistent decline in prices across various sectors. This deflationary environment posed challenges for the Japanese economy, as falling prices can discourage consumer spending and investment. Policymakers closely monitored the CPI numbers to assess the effectiveness of their economic policies and implement measures to combat deflation.
4. Impact of COVID-19 on CPI: The COVID-19 pandemic had a significant impact on consumer prices worldwide. During the initial stages of the pandemic, there were disruptions in the global supply chains, which led to supply shortages and price increases for certain goods. However, as lockdown measures were implemented and consumer spending declined, the demand for many goods and services decreased, resulting in deflationary pressures. CPI numbers reflected these changes, showing fluctuations in prices for essential items like food, healthcare services, and transportation.

These case studies illustrate how CPI numbers can provide valuable insights into economic trends, such as recessions, hyperinflation, deflation, and the impact of external factors like pandemics on consumer prices. Monitoring and analyzing CPI data are essential for policymakers, economists, and businesses to make informed decisions and understand the overall health of an economy (Fig. 2.6).

2.7.3 The Weighting Pattern for 2019-Based CPI for General Households

The CPI weights reflect the relative importance of each good or service in the basket. The Department of Statistics Singapore has showcased the weighting pattern for the 2019-based CPI which was derived from the expenditure values obtained from the Household Expenditure Survey (H.E.S.) conducted between October 2017 and September 2018 and updated to 2019 values by taking into account price changes between 2017/18 and 2019. The weights for CPI are

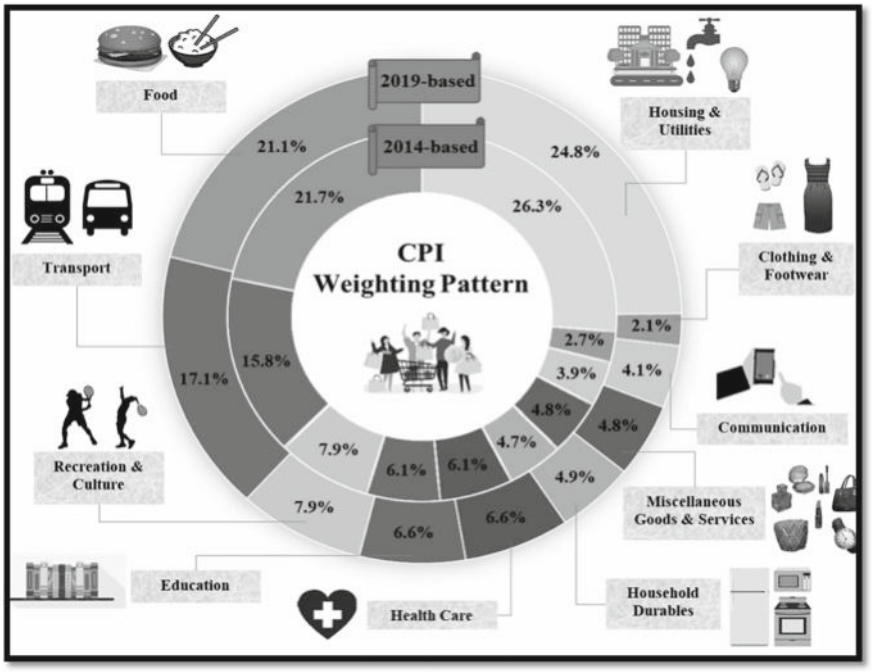


Fig. 2.6 CPI weighting pattern

as follows: https://www.singstat.gov.sg/find-data/search-by-theme/economy/prices-and-price-indices/related-info/faq-on-cpi.

Example: An enquiry into the budgets of middle-class families of a certain town revealed that on an average the percent expenses on different groups of expenditure were as follows. Food 40, rent 20, clothing 12, fuel and light 10, and miscellaneous 18. The group index numbers for the current year as compared with a fixed base period were respectively 380, 170, 325, 230, and 270. Calculate the Consumer Price Index number for the current year (Table 2.8).

Table 2.8 Calculating CPI using weights and index numbers

Groups	Weight (W)	Index number (I)	$I * W$
Food	40	380	15,200
Rent	20	170	3400
Clothing	12	325	3900
Fuel and light	10	230	2300
Miscellaneous	18	270	4860
Total	100		29,660

Solution

Consumer Price Index number is given as CPI $= \frac{\sum I*W}{\sum W} = 29660/100 = 296.60$.

2.7.4 Calculation of CPI

In most of the scenarios, CPI is considered as "standard of living." We must be very watchful of the two related terms. Though cost is one of the important factors for judging the standard of living of people, there are several other aspects which are sometimes ignored. These can be size of the family, region of study, education and employment, and many more.

The Sixth International Conference of Labor Statistics held under I.L.O. in 1949 recommended that "Cost of Living index number should be appropriately renamed as consumer price index numbers or retail price index numbers."

There are two types of CPI calculation methods:

- Family budget method.
- Aggregate expenditure method.

Aggregate expenditure method: In this method, we consider quantity consumed to be as weights. Hence, we multiply the current year price with base year quantity and take the sum aggregate. This value is again divided by the aggregate expenditure of that commodity in the base year to obtain the final value that represents CPI.

Family budget method as the name suggests, is a concept, wherein family budgets are studied for a large section of population. The consumption is considered as weight here across variety of consumer goods. This is an example of weighted average of price relative method $\frac{\sum WP}{W}$.

Consider this example in which we need to compute CPI for 1982 with base period being 1975. Let's solve using both the methods discussed (Tables 2.9 and 2.10).

Solution

Aggregate Expenditure Method:

$$\text{CPI} = \frac{\Sigma p1q0}{\Sigma p0q0} \times 100 = \frac{297.8}{214.31} \times 100 = 138.95$$

Table 2.9 Consumer goods with quantity and prices of year 1975 and 1982

Year	Commodity	Wheat	Urad Dal	Milk	Onion	Sugar	Tea	Kerosene
1975	Quantity	1.57	3.17	1.5	0.5	4	17	1.12
	Price	58	5	30	15	6	0.5	20
1982	Price	1.95	1.45	3	0.75	4.7	22	1.85

Table 2.10 Calculation of aggregate expenditure method and family budget method for various commodities

Commodity	P_0	q_0	P_1	Weights p_0q_0	p_1q_0	$P = \frac{P_1}{P_0} * 100$	W^*P
Wheat	1.57	58	1.95	91.06	113.1	124.20	11,310.00
Urad Dal	3.17	5	1.45	15.85	7.25	45.24	725.00
Milk	1.5	30	3	45	90	200.00	9000.00
Onion	0.5	15	0.75	7.5	11.25	150.00	1125.00
Sugar	4	6	4.7	24	28.2	117.50	2820.00
Tea	17	0.5	22	8.5	11	129.41	1100.00
Kerosene	1.12	20	1.85	22.4	37	165.18	3700.00
Total				214.31	297.8	932.04	29,779.99

Family Budget Method:

$$\text{CPI} = \frac{\Sigma WP}{\Sigma W} = \frac{29779.99}{214.31} = 139.05.$$

2.8 Wholesale Price Index Number (WPI)

WPI is the economic barometer that indicates the inflation levels in the economy with respect to wholesale prices of commodities. From the production units, throughout the supply chain of distribution, there is a price that is added at each level. Wholesale price is the amount that is added at a stage where the finished product from the warehouse is distributed to various retail stores/local supermarkets, etc. (Fig. 2.7).

Initially there were a total of 112 commodities that were categorized into six major groups by the economic advisor. Later the Standard International Trade Classification

Fig. 2.7 WPI

comprised 555 individual quotations. Quotation of prices was collected through official sources like State Bank of India, Trade associations, Chamber of Commerce, etc. (Table 2.11).

Over a period of time, index numbers of wholesale prices in India, revised series was published weekly by the Economic Adviser using weighted arithmetic mean. These weekly index numbers were then averaged to monthly index numbers (Table 2.12).

The latest announcement by the economic advisor about the wholesale prices in the country:

From the official website, consider the below tabular information from the Economic Survey, Government of India, 2004–2005. Calculate and increase/decrease in the CPI and WPI to understand the economy in general and comment on the solution thus obtained (Table 2.13).

The calculation for industrial workers (Fig. 2.8):

Table 2.11 Index numbers of wholesale prices of major groups and sub-groups

Index number of wholesale prices				
By major groups and sub-groups				
Major group/sub-groups	1994–95	1999–00	2000–01	2001–02
I. Primary articles	115.8	158.0	162.5	168.4
1. Food articles	112.8	165.5	170.5	176.1
2. Non-food articles	124.2	143.0	146.5	152.9
3. Minerals	104.9	110.4	113.5	119.3
II. Fuel, power, light, and lubricants	108.9	162.0	208.1	226.7
III. Manufactured products	112.3	137.2	141.7	144.3
1. Food products	114.1	151.3	145.7	145.4
2. Beverages, tobacco, and tobacco products	118.3	174.1	179.8	193.8
3. Textiles	118.2	115.0	119.9	119.3
4. Wood and wood products	110.9	193.9	180.0	174.4
5. Paper and paper products	106.1	149.6	165.4	172.8
6. Leather and leather products	109.7	154.6	149.6	141.0
7. Rubber and plastic products	106.4	123.6	125.5	126.0
8. Chemicals and chemical products	116.6	155.2	164.4	169.0
9. Non-metallic mineral products	110.9	127.4	133.9	144.0
10. Basic metals, alloys, and metal products	108.4	135.0	140.3	140.7
11. Machinery and machine tools including electricity machinery	106.0	116.1	123.0	129.1
12. Transport equipments and parts	107.4	135.4	143.4	146.8
All commodities	**112.6**	**145.3**	**155.7**	**161.3**

Source Office of the economic adviser, Ministry of Industry
(Statistical Pocket Book India 2002, Page 207), Base: 1993–94 =100

Table 2.12 Index numbers of wholesale prices (base 1993–94 =100)

	Index numbers of wholesale prices (Base: 1993–94 =100)												
	Month	April	May	June	July	August	September	October	November	December	January	February	March
All commodities (100)	2001–02	159.9	160.3	160.8	161 1	161.7	161.7	162.5	162.3	161.8	161	160.8	161.9
	2002–03	162.3	162.8	164.7	165.6	167.1	167.4	167.5	167.8	167.2	167.8	169.4	171.6
	2003–04	173.1	173.4	173.5	173.4	173.7	–	–	–	–	–	–	–
Primary articles (22.03)	2001–02	165.1	167.1	169.7	168.6	169.4	170.2	170.8	170	160.3	166.4	166.7	167.3
	2002–03	169	168.9	171.9	172.3	175.4	176,1	175.2	176.5	174.7	173.8	176.5	171.9
	2003–04	179.9	180.8	183.8	180.5	178.6	–	–	–	–	–	–	–
Fuel power, light, and lubricants (14.23)	2001–02	222 7	222.6	222.5	222.3	226.1	226.3	2304	230.6	229	228.1	227.3	233.9
	2002–03	230.4	230 4	233 8	238 4	237 8	2388	2409	2408	2388	241 8	244 3	254.1
	2003–04	254 2	247 6	246.1	249 3	249 5	–	–	–	–	–	–	–
Manufactured products (63.75)	2001–02	114.2	144.1	144	144.9	144.6	144.3	144.4	144.4	144.2	144.2	143.9	144.1
	2002–03	144.9	145.5	146.8	147.1	148.5	145.6	1484	145.5	148.6	149.2	150.3	151
	2003–04	152.6	154.3	153.8	154.1	155.1	–	–	–	–	–	–	–

Source CSO, Monthly Abstract of Statistics, October 2003

Table 2.13 CPI and WPI of workers

Year	CPI of industrial workers (1982 = 100)	CPI of urban non-manual employees	CPI agricultural laborers (1986–87 = 100)	WPI (1993–94 = 100)
1995–96	313	257	234	121.6
1996–97	342	283	256	127.2
1997–98	366	302	264	132.8
1998- 99	414	337	293	140.7
1999–00	428	352	306	145.7
2000–01	444	352	306	155.7
2001–02	463	390	309	161.3
2002–03	482	405	319	166.8
2003–04	500	420	331	175.9

Economic survey, Government of India, 2004–05

$$\text{Year } 1996-97 = \frac{342-313}{313} * 100 = 9.26\% \text{ and so on.}$$

The calculation for agricultural laborers:

GOVERNMENT OF INDIA
MINISTRY OF COMMERCE & INDUSTRY
DEPARTMENT FOR PROMOTION OF INDUSTRY AND INTERNAL TRADE
OFFICE OF THE ECONOMIC ADVISER

New Delhi, Date: January 16th, 2023

PRESS RELEASE

Index Numbers of Wholesale Price in India for the Month of December, 2022 (Base Year: 2011-12)

The annual rate of inflation based on all India Wholesale Price Index (WPI) number is 4.95% (Provisional) for the month of December, 2022 (over December, 2021) against 5.85% recorded in November, 2022. Decline in the rate of inflation in December, 2022 is primarily contributed by fall in prices of food articles, mineral oils, crude petroleum & natural gas, food products, textiles and chemicals & chemical products. The index numbers and inflation rate for the last three months of all commodities and WPI components are given below:

Index Numbers & Annual Rate of Inflation (Y-o-Y in %)*							
All Commodities/Major Groups	Weight (%)	Oct-22 (F)		Nov-22 (P)		Dec-22 (P)	
		Index	Inflation	Index	Inflation	Index	Inflation
All Commodities	100.0	152.9	8.67	152.1	5.85	150.4	4.95
I. Primary Articles	22.6	181.2	11.17	177.7	5.52	172.4	2.38
II. Fuel & Power	13.2	158.0	25.40	159.6	17.35	158.0	18.09
III. Manufactured Products	64.2	141.9	4.42	141.5	3.59	141.1	3.37
Food Index	24.4	177.7	6.60	174.3	2.17	170.3	0.65

*Note: P: Provisional, F: Final, *Annual rate of WPI inflation calculated over the corresponding month of previous year*

Fig. 2.8 Index numbers of wholesale prices for the month of December 2022 (base year 2011–12)

$$\text{Year } 1996-97 = \frac{256-234}{234} * 100 = 9.40\% \text{ and so on.}$$

Similar calculations are done for Wholesale Price Index (W.P.I.) (Table 2.14).

The datasets are comparable as they are CPI values of different divisions of laborers, and also, they have a fixed base year. Many fluctuations can be noticed in each category of workers; none of them depicts sustainable progress.

2.8.1 Some Case Studies on WPI

1. Indian WPI and Inflation: India's Wholesale Price Index (WPI) is a measure of the average change in the selling prices received by domestic producers for their goods and services at the wholesale level. In the past, WPI numbers in India have been used as an indicator of inflation. A case study could focus on a period of high inflation in India, such as during 2010–2011. The WPI numbers during this period showed a significant increase in prices, driven by factors such as rising crude oil prices, food inflation, and global commodity price fluctuations. These WPI numbers were crucial for policymakers in assessing the inflationary pressures on the economy and formulating appropriate monetary and fiscal policies to manage inflation.
2. Impact of GST Implementation in India: The Goods and Services Tax (GST) is a comprehensive indirect tax implemented in India in 2017. It replaced multiple taxes levied by the central and state governments. The implementation of GST had a significant impact on WPI numbers in various sectors. A case study could focus on specific industries like manufacturing or services and examine how the WPI numbers changed after the introduction of GST. The study could highlight the impact of tax rate changes, input tax credits, and the overall ease of doing business on the WPI numbers in different sectors.
3. WPI and Supply Chain Disruptions: Supply chain disruptions, such as natural disasters or geopolitical events, can have an impact on WPI numbers. For example, a case study could analyze the WPI numbers during the 2011 earthquake and tsunami in Japan. The disruption caused by the disaster affected industries like automobile manufacturing, electronics, and machinery. The WPI numbers reflected the increase in prices of affected goods due to supply shortages and disruptions in production. This case study would demonstrate how WPI numbers can capture the impact of supply chain disruptions on wholesale prices and provide insights into the economic consequences of such events.
4. WPI and Agricultural Price Fluctuations: Agricultural products play a significant role in WPI calculations. Fluctuations in agricultural prices can impact the overall WPI numbers. A case study could examine the WPI numbers during a period of agricultural price volatility, such as a drought or a bumper crop year. The study could focus on specific agricultural commodities and analyze how their price

Table 2.14 CPI of industrial workers, urban non-manual employees and agricultural labourers along with WPI (1993-94 =100)

Year	CPI of industrial workers (1982 =100)	CPI of urban non-manual employees	CPI agricultural laborers(1986–87 = 100)	WPI (1993–94 =100)
1995–95				
1995–97	↓9.26%	↑9.19%	↑9.40%	↓4.60%
1997–98	↓7.01%	↓6.29%	↓3.13%	↓4.40%
1998–99	↑13.11%	↑10.39%	↑10.98%	↓5.94%
1999–00	↑11.59%	↓4.26%	↓4.43%	↓3.26%
2000–01	↓3.73%	↓0.00%	↓0%	↓7.15%
2001–02	↓4.27%	↑9.54%	↓0.98%	↓29.38%
2002–03	↓3.93%	↓3.70%	↓3.23%	↓3.40%
2003–04	↓3.73%	↓3.54%	↓3.76%	↓5.17%

changes influenced the WPI. This analysis would help understand the relationship between agricultural production, market dynamics, and the WPI.

These case studies highlight the importance of WPI numbers in assessing inflationary pressures, understanding the impact of policy changes, supply chain disruptions, and agricultural price fluctuations. The WPI provides valuable information for policymakers, businesses, and economists to make informed decisions related to pricing, production, and economic policies.

2.9 Tax Price Index Numbers-TPI

Tax price index numbers refer to indices that measure changes in tax rates or tax burdens over time. In simpler terms it is an answer to the question, by how much should the consumer income rise in order to match the existing purchasing power in an economy. Let us consider that John's annual disposable income is 60,000$. In the following year, the inflation and tax rates are hiked by 2.5 and 3.25%, respectively. In order to maintain the same level of purchasing power, his disposable income is expected to rise by 5.75%. This way, T.P.I. is a more sophisticated index number that accounts for many factors such as retail prices and direct taxes. While tax price index numbers are less commonly used compared to other types of index numbers, here is a case study that showcases their potential application.

2.9.1 Case Study: Tax Reform and Tax Price Index Numbers

In 2017, the United States implemented comprehensive tax reform, known as the Tax Cuts and Jobs Act (TCJA). The TCJA aimed to simplify the tax system, promote economic growth, and reduce the tax burden on businesses and individuals. Tax price index numbers could be utilized to analyze the impact of this tax reform on taxpayers and the overall economy (Fig. 2.9).

1. Tax Burden Analysis: Tax price index numbers can help assess the changes in the tax burden on different income groups and industries. By comparing tax price index numbers before and after the tax reform, policymakers and economists can evaluate how the tax burden shifted across various income levels. This analysis would provide insights into the distributional effects of the tax reform and its impact on income inequality.
2. Business Competitiveness: Tax price index numbers can be used to gauge the impact of tax reforms on business competitiveness. Lower tax rates or changes in tax structures can affect businesses' investment decisions, profitability, and international competitiveness. By tracking tax price index numbers, policymakers can monitor changes in tax burdens and assess how these changes influence business decisions, job creation, and economic growth.

Fig. 2.9 Tax and price index numbers (Jan 1987 = 100)

3. Behavioral Effects: Tax price index numbers can shed light on the behavioral changes induced by tax reforms. For example, if tax rates on certain goods or services are reduced, tax price index numbers can indicate whether there is a shift in consumption patterns or increased economic activity in specific sectors. This information can be useful for policymakers to understand the response of taxpayers to tax incentives and tailor policies accordingly.

It is important to note that tax price index numbers may face challenges due to complexities in tax systems, including multiple tax rates, exemptions, and deductions. The calculation of tax price indices requires accurate and comprehensive tax data, which may not always be readily available. Nonetheless, by utilizing tax price index numbers, policymakers can evaluate the effects of tax reforms, make data-driven decisions, and assess the overall impact on taxpayers and the economy.

T.P.I. is published by Office of National Statistics by setting the base value of January 1987 as 100. In January of 2017, the rate of inflation as measured by the index rose 3.1% over the previous 12 months. This number is relatively low, historically speaking. For example, the TPI reflected a 25.5 year-over-year change in January 1975, reflecting the need for incomes to rise 25.5% over the 12-month for a person to maintain the same purchasing power and quality of life.

2.10 Crime Index Numbers

Crime index numbers are composite statistical measures that provide a quantitative assessment of the overall crime situation in a particular region or jurisdiction. They provide a comprehensive and quantitative understanding of crime rates, patterns, and trends within a given area or jurisdiction. These indices help policymakers, law enforcement agencies, researchers, and communities make informed decisions, allocate resources effectively, and develop targeted crime prevention strategies. By

monitoring crime index numbers over time, stakeholders can assess the impact of interventions, evaluate public safety, and work toward fostering safer societies.

These indices are typically derived from a combination of crime-related data, such as reported offenses, arrests, convictions, and other relevant information collected by law enforcement agencies. By aggregating these data points, crime index numbers offer a summarized view of the prevalence and severity of different types of crimes within a given area.

2.10.1 Components of Crime Index Numbers

1. Reported Crimes: One of the primary components of crime index numbers is the count of reported crimes. This includes offenses reported to law enforcement agencies by victims or witnesses, encompassing various categories such as violent crimes (e.g., homicide, assault), property crimes (e.g., burglary, theft), and other types of criminal activities.
2. Clearance Rates: Clearance rates indicate the proportion of reported crimes that have been resolved or closed by law enforcement agencies. They reflect the effectiveness of investigations and the ability to identify and apprehend suspects. Higher clearance rates may indicate more successful law enforcement efforts and serve as an indicator of public safety.
3. Arrests and Convictions: The number of arrests and convictions resulting from reported crimes is another key component of crime index numbers. It signifies the criminal justice system's ability to apprehend perpetrators, bring them to trial, and secure convictions. Tracking the rate of arrests and convictions helps gauge the effectiveness of law enforcement and judicial processes in combating crime.
4. Crime Rates: Crime rates, including violent crime rates and property crime rates, are vital components of crime index numbers. These rates are calculated by dividing the number of reported crimes by the population size and are often expressed per 1000 or 100,000 residents. Crime rates provide a standardized measure that facilitates comparisons across different jurisdictions and time periods.

2.10.2 Significance of Crime Index Numbers

1. Crime Prevention and Resource Allocation: Crime index numbers play a crucial role in formulating targeted crime prevention strategies. By analyzing crime patterns and trends, policymakers and law enforcement agencies can identify high-crime areas and allocate resources accordingly. These indices help prioritize interventions, such as increased patrols, community policing initiatives, or enhanced security measures, to address specific crime hotspots effectively.

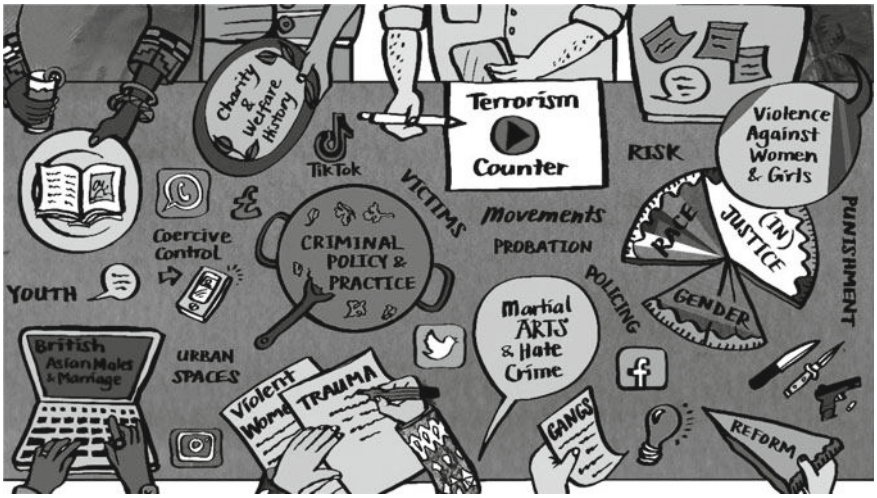
2. Assessing Public Safety: Crime index numbers provide a benchmark for assessing public safety and evaluating the success of crime reduction efforts. A declining crime index may indicate improved safety and suggest that implemented strategies are yielding positive results. Conversely, an increasing crime index can signal the need for additional interventions and policy adjustments to ensure public safety.
3. Research and Evaluation: Crime index numbers serve as valuable data sources for researchers studying crime patterns, risk factors, and the effectiveness of crime prevention programs. By analyzing long-term trends and correlating crime indices with various socioeconomic variables, researchers can gain insights into the underlying causes of crime and identify evidence-based strategies to reduce criminal activities.
4. Community Engagement: Crime index numbers foster community engagement and collaboration in crime prevention efforts. By disseminating crime statistics and indices to the public, communities can gain awareness of local crime issues, encouraging citizens to adopt preventive measures and work alongside law enforcement agencies. Community involvement and vigilance are crucial in reducing crime and improving overall public safety.

Neighborhood Scout's Crime Index is a value ranging between 0 and 100, where 100 is referred to as the safest score. A crime index of 75 means that the neighborhood is 75% safer than the other neighborhoods of America.

F.B.I. or Federal Bureau of Investigation is responsible for compiling all the crime statistics on an annual basis and presenting the crime index or the National Uniform Crime Report. Similar to the idea of a "basket of goods" in the calculation of C.P.I., the F.B.I. has also consolidated all crimes into major eight crimes divided into two branches, namely *violent crime* and *property related crime.* These major categories are:

1. Murder.
2. Forcible Rape.
3. Larceny-theft
4. Burglary.
5. Motor Vehicle Theft.
6. Aggravated Assault.
7. Arson.
8. Robbery.

There was always a need for a meaningful weighted index number that comprehends all these crime data and incorporates a legitimate measure of harm/pain/loss as weights. The resulted index number forms a base for the penalty or punishment that the accused/offender deserves.



The C.C.H.I.—Cambridge Crime Harm Index—May 2020, is one among the latest research in the field of crime that uses index numbers to measure the harm that each type of offense causes to the victims, based on which a baseline punishment can be ascertained and ordered to the accused in the court of justice C.C.H.I. was proposed by Lawrence Sherman, Peter William Neyroud and Eleanor Neyroud in 2016.

This index is not just useful in criminal justice but is also used as an important reference code by ATHENA—System of Criminal Recording and Management System used by various forces in the United Kingdom to improvise and ease the addition of CCHI scores to the available police data.

2.11 Environmental Quality Index

Environmental Quality Index: The Environmental Quality Index is a measure of the environmental health and sustainability of a region. It considers various environmental indicators such as air and water qualities, waste management, biodiversity, and natural resource conservation. The index helps gauge the overall environmental performance and highlights areas that need improvement.

Uses and Applications:

- Assessing the environmental impact of policies and projects.
- Monitoring progress toward environmental goals and targets.
- Comparing the environmental performance of different countries or regions.
- Raising awareness about environmental issues and encouraging sustainable practices.

Example: To calculate the Environmental Quality Index (EQI) for a particular region based on different environmental parameters. For this example, let us consider a simplified EQI that takes into account four environmental parameters: air quality, water quality, biodiversity, and waste management.

Step 1: Define the Environmental Parameters

Let us define the ideal ranges for each parameter:

- Air Quality: Ideally Air Quality Index (AQI) is below 50, indicating good air quality.
- Water Quality: Ideally, the Water Quality Index should be above 80, indicating good water quality.
- Biodiversity: Ideally, a Biodiversity Index is above 75, indicating high biodiversity.
- Waste Management: Ideally, a Waste Management Index is above 70, indicating effective waste management practices.

Step 2: Collect Data

Let us assume we have data for the region as follows:

- Air Quality (AQI): 65.
- Water Quality Index: 75.
- Biodiversity Index: 80.
- Waste Management Index: 68

Step 3: Calculate Scores for Each Parameter

To calculate the scores for each parameter, we will use a simple linear scoring system where higher values result in higher scores (0–100). For parameters within the ideal range, the score will be 100, and for values outside the ideal range, the score will decrease linearly.

(a) Air Quality Score:

$$\text{Air Quality Score} = 100 - (\text{abs}(\text{ideal\_ AQI} - \text{actual AQI}) * 2)$$
$$\text{Air Quality Score} = 100 - (\text{abs}(50 - 65) * 2)$$
$$\text{Air Quality Score} = 100 - (15 * 2)$$
$$\text{Air Quality Score} = 100 - 30$$
$$\text{Air Quality Score} = 70$$

(b) Water Quality Score:

$$\text{Water Quality Score} = 100 - (\text{abs}(\text{ideal\_Water\_quality} - \text{actual\_water\_quality}) * 1.25)$$
$$\text{Water Quality Score} = 100 - (\text{abs}(80 - 75) * 1.25)$$
$$\text{Water Quality Score} = 100 - (5 * 1.25)$$
$$\text{Water Quality Score} = 100 - 6.25$$
$$\text{Water Quality Score} = 93.75$$

(c) Biodiversity Score:

$$\text{Bio diversity Score} = 100 - (\text{abs}(\text{ideal\_ biodiversity} - \text{actual\_ biodiversity}) * 1.33)$$
$$\text{Bio diversity Score} = 100 - (\text{abs}(75 - 80) * 1.33)$$
$$\text{Bio diversity Score} = 100 - (5 * 1.33)$$
$$\text{Bio diversity Score} = 100 - 6.65$$
$$\text{Bio diversity Score} = 93.35$$

(d) Waste Management Score:

$$\text{Waste Management Score} = 100 - (\text{abs}(\text{ideal\_ waste\_ management} - \text{actual\_ waste\_ management}) * 2)$$
$$\text{Waste Management Score} = 100 - (\text{abs}(70 - 68) * 2)$$
$$\text{Waste Management Score} = 100 - (2 * 2)$$
$$\text{Waste Management Score} = 100 - 4$$
$$\text{Waste Management Score} = 96$$

Step 4: Calculate the Environmental Quality Index (EQI)

To calculate the overall EQI, we take the average of the scores obtained for each parameter.

$$\text{Environmental Quality Index(EQI)} = (\text{Air Quality Score} + \text{Water Quality Score} + \text{Biodiversity Score} + \text{Waste Management Score})/4$$
$$\text{Environmental Quality Index(EQI)} = (70 + 93.75 + 93.35 + 96)/4$$
$$\text{Environmental Quality Index(EQI)} = 353.1/4$$
$$\text{Environmental Quality Index(EQI)} \approx 88.275$$

Step 5: Interpretation

The calculated Environmental Quality Index (EQI) is approximately 88.275 out of 100. This indicates that the overall environmental quality in the region is relatively good, with minor room for improvement in some parameters, particularly air quality. However, it is essential to conduct a comprehensive analysis and consider other relevant environmental factors to have a complete understanding of the region's environmental health and potential areas of concern. Additionally, environmental policies and initiatives can be formulated based on the EQI to improve the region's environmental sustainability.

2.12 The Health Index

The Health Index is a composite measure that evaluates the overall health status and healthcare system performance of a population. It incorporates indicators such as life expectancy, mortality rates, disease prevalence, access to healthcare services, and public health initiatives.

Uses and Applications:

- Benchmarking the healthcare system against international standards.
- Identifying health disparities and areas requiring targeted interventions.
- Assessing the effectiveness of public health programs and policies.
- Informing healthcare resource allocation and planning.

Example: Consider this data related to calculating a Health Index for an individual based on different health parameters. For this example, let us consider a simplified Health Index that takes into account three health parameters: Body Mass Index (BMI), Blood Pressure (BP), and Cholesterol levels.

Step 1: Define the Health Parameters

Let us define the ideal ranges for each parameter:

– BMI: Ideally between 18.5 and 24.9.
– Blood Pressure (BP): Ideally below 120/80 mmHg.
– Cholesterol: Ideally total Cholesterol below 200 mg/dL.

Step 2: Collect Data

Let us assume we have data for an individual as follows:

– BMI: 26.5.
– Blood Pressure (BP): 130/85 mmHg.
– Cholesterol: 210 mg/dL.

Step 3: Calculate Scores for Each Parameter

To calculate the scores for each parameter, we will use a simple linear scoring system where higher values result in lower scores (0–100). For parameters within the ideal range, the score will be 100, and for values outside the ideal range, the score will decrease linearly.

(a) BMI Score:

$$\text{BMI score} = 100 - (\text{abs}(\text{ideal\_ BMI} - \text{individual\_ BMI}) * 4)$$
$$\text{BMI score} = 100 - (\text{abs}(22.2 - 26.5) * 4)$$
$$\text{BMI score} = 100 - (4.3 * 4)$$
$$\text{BMI score} = 100 - 17.2$$
$$\text{BMI score} = 82.8$$

(b) Blood Pressure (BP) Score:

For Blood Pressure, we will calculate the average of systolic and diastolic scores.

Systolic Score:

$$\begin{aligned}
&\text{Systolic score} = 100 - (\text{abs}(\text{ideal_ systolic} - \text{Individual_ systolic}) * 0.5) \\
&\text{Systolic score} = 100 - (\text{abs}(120 - 130) * 0.5) \\
&\text{Systolic score} = 100 - (10 * 0.5) \\
&\text{Systolic score} = 100 - 5 \\
&\text{Systolic score} = 95
\end{aligned}$$

Diastolic Score:

$$\begin{aligned}
&\text{Diastolic score} = 100 - \ (\text{abs}\,(\text{ideal_diastolic} - \text{individual_diastolic}\,) * 0.5\,) \\
&\text{Diastolic score} = 100 - \ (\text{abs}(80 - 85\,) * 0.5\,) \\
&\text{Diastolic score} = 100 - (5 * 0.5) \\
&\text{Diastolic score} = 100 - 2.5 \\
&\text{Diastolic score} = 97.5
\end{aligned}$$

BP Score (average of systolic and diastolic scores):

$$\begin{aligned}
&\text{BP Score} = (\text{Systolic Score} + \text{Diastolic Score}\,)/2 \\
&\text{BP Score} = (95 + 97.5)/2 \\
&\text{BP Score} = 192.5/2 \\
&\text{BP Score} = 96.25
\end{aligned}$$

(c) Cholesterol Score:

$$\begin{aligned}
&\text{Cholesterol score} = 100 - \ (\text{abs}\,(\text{ideal_ Cholesterol} - \text{individual_ Cholesterol}\,) * 0.5\,) \\
&\text{Cholesterol score} = 100 - (\text{abs}(200 - 210) * 0.5) \\
&\text{Cholesterol score} = 100 - (10 * 0.5) \\
&\text{Cholesterol score} = 100 - 5 \\
&\text{Cholesterol score} = 95
\end{aligned}$$

Step 4: Calculate the Health Index

To calculate the overall Health Index, we take the average of the scores obtained for each parameter.

$$\text{Health Index} = (\text{BMI Score} + \text{BP Score} + \text{Cholesterol Score}\ )/3$$
$$\text{Health Index} = (82.8 + 96.25 + 95\ )/3$$
$$\text{Health Index} = 273.05/3$$

Step 5: Interpretation

The calculated Health Index is approximately 91.02 out of 100. This indicates that the individual's health is generally good, with minor room for improvement in some health parameters (specifically BMI and Blood Pressure). It is essential to consult a healthcare professional for personalized advice and recommendations on maintaining and improving overall health.

2.13 Education Index Number

"Education is not the learning of facts, but the training of the mind to think."—Albert Einstein.

The Education Index is a statistical measure that assesses the education level and literacy rate of a population. It is typically used to compare the education performance of different countries or regions. The index is often constructed using indicators such as literacy rates, enrollment ratios, and educational attainment levels. A higher Education Index value indicates a higher level of education and human capital development in a particular area.

The Education Index is one component under the Human Development Index (H.D.I.) which is a key value depicting the performance of the states upon evaluating all the factors of an education system. https://www.insightsonindia.com/2021/06/07/education-index-ranking/.

Performance Grading Index, P.G.I., is a tool used by the Ministry of Education to evaluate all the indicators of education in every state of the country. The latest scores of P.G.I. (2019–2020) are depicted in Fig. 2.10.

Andaman and Nicobar, Punjab, Chandigarh, Kerala, and Tamil Nadu are the states in the A++ category scoring over 900 out of 1000 points, whereas states such as Bihar and Meghalaya scored the least mainly due to the lack of infrastructure and educational facilities.

$$\text{Education Index (E.I.)} = \frac{\text{Mean Years of Schooling Index (MYSI)}}{\text{Expected Years of Schooling Index (EYSI)}}$$

$$\text{Mean Years of Schooling Index (MYSI)} = \frac{\text{Mean Years of Schooling}}{15}$$

MYSI is the average number of completed years of education for a population that is 25 years and above. The index is divided by 15 which is forecasted as the maximum indicator for 2025.

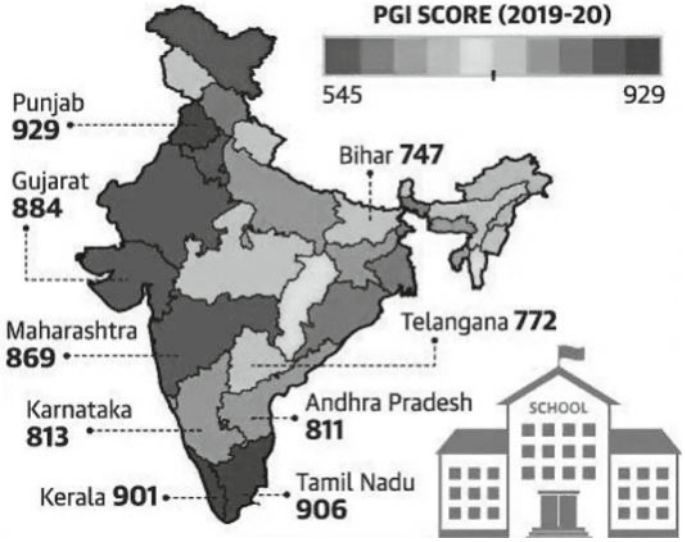


Fig. 2.10 Performance grading index of India

$$\text{Expected Years of Schooling Index (EYSI)} = \frac{\text{Expected Years of Schooling}}{18}$$

EYSI for a child who is of age to start school is defined as, what is the expected number of schooling years that the child would receive if the age-specific enrolments rates prevail throughout the child's life. The index is divided by 18, which is the general age at which students are awarded a master's degree.

For example, in a region where the expected year of schooling is given as 24 and the mean years of schooling is 26, the Education Index would be

$$\text{Education Index (E.I.)} = \frac{\frac{\text{Mean Years of Schooling Index (MYSI)}}{15}}{\frac{\text{Expected Years of Schooling Index (EYSI)}}{18}}$$

$$= \frac{(26/15)}{(24/18)} = \frac{1.733}{1.333} = 1.3$$

These various index numbers provide valuable data and insights into different aspects of society and can be instrumental in decision-making, policy formulation, and addressing various challenges. As with any index, the construction and selection of indicators should be carefully considered to ensure the index's relevance and accuracy in capturing the intended information.

2.14 Types of Index Number Based on Weights and Formula

2.14.1 Laspeyre's Index—Output Inflator

Named after a German economist Étienne Laspeyre, this method of index calculation was formulated in 1871. His rationale was to choose base year values as weights. Therefore, this method is also called base weighted index which was applicable at times when the effect of price rise had to be reduced (Figs. 2.11 and 2.12).

$$\text{Laspeyre's index} = \frac{\sum \left(X_{c,t_n}\right) * \sum \left(X'_{c,t_0}\right)}{\sum \left(X_{c,t_0}\right) * \sum \left(X'_{c,t_0}\right)} * 100$$

where X' is the weights for calculating the index number of X. (c, t_n) refers to cost at time t_n and (c, t_0) refers to cost in the base year.

$$\text{Laspeyre's price index} = P_{01}^{L} = \frac{\sum p_1 q_0}{\sum p_0 q_0} * 100$$

In price index calculation the reason to opt for Laspeyre's index was to reduce the price rise to adjust the inflation. But this idea became a setback as it was a condition of upward bias. This method of index calculation helps you understand, how much of the total price increase is due to the variation in the quantity levels and how much is due to inflation. In case we need to compute the quantity index number (Q_{01}^{L}), then the base year price of the commodity will be the weights. The formula is thus, $Q_{01}^{L} = \frac{\sum q_1 p_0}{\sum q_0 p_0} * 100$.

2.14.2 Paasche's Index—Output Deflator

In 1875, Hermann Paasche, a German economist believed in adding current year values as weights. This gives a better reflection of the consumption pattern. Therefore, this method is also called the current weighted index. But this method of index calculation understates the inflation therefore Paasche's index method was not applicable/effective in all fields of study. The formula to calculate the price index number given by Paasche is (Fig. 2.13)

$$\text{Paasche's index} = \frac{\sum \left(X_{c,t_n}\right) * \sum \left(x'_{c,t_n}\right)}{\sum \left(X_{c,t_0}\right) * \sum \left(X_{c,t_n}\right)} * 100$$

where X' is the weights for calculating the index number of X. (c, t_n) refers to cost at time t_n and (c, t_0) refers to cost in the base year.

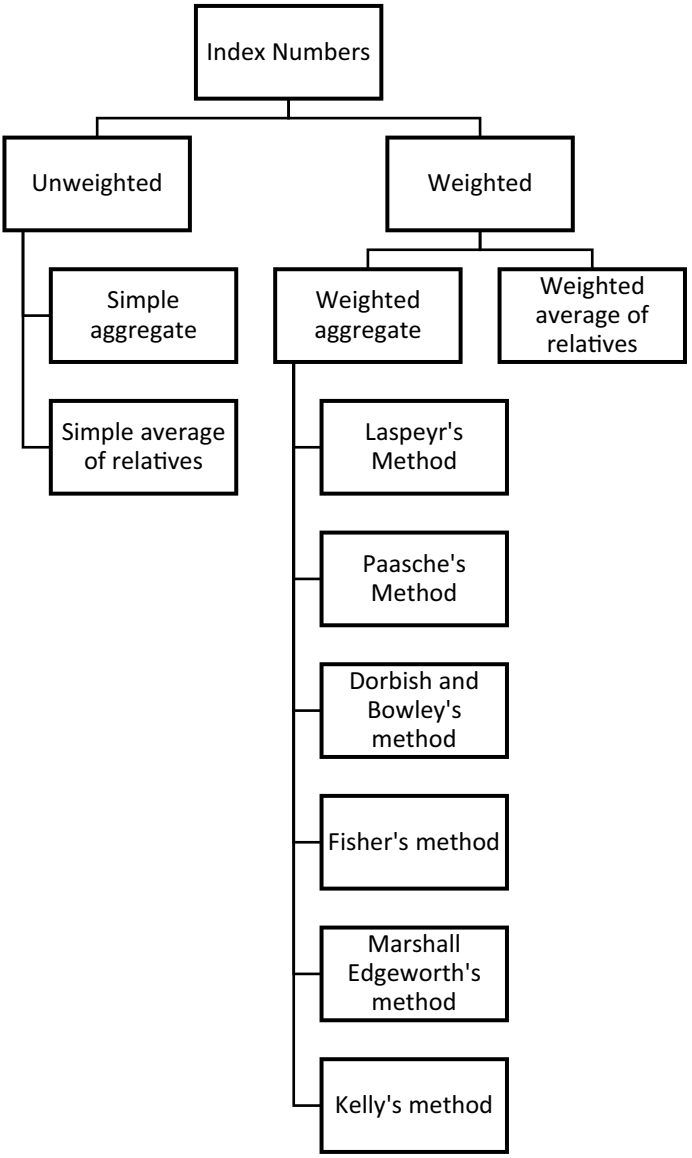


Fig. 2.11 Classification of index numbers with respect to their calculations

$$\text{Paasche's price index } = P_{01}^{P} = \frac{\sum p_1 q_1}{\sum p_0 q_1} * 100$$

Interpretation: This formula helps you decipher the difference in price in today's basket of goods in today's dollars versus the price for the same basket of goods in

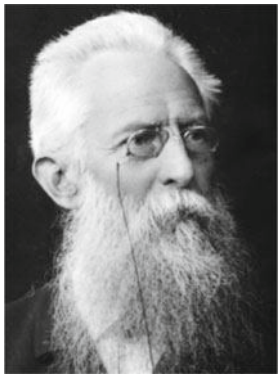


Fig. 2.12 Étienne Laspeyre's

Fig. 2.13 Hermann Paasche

base year dollars? Analysts who need to measure the real value of output for a specific period will find this measure useful. In case we need to compute the quantity index number (Q_{01}^{P}), then the current year price of the commodity will be the weight. The formula is thus, $Q_{01}^{L} = \frac{\sum q_1 p_1}{\sum q_0 p_1} * 100$ [11]. The picture above shows the comparison between both Laspeyre's and Paasche's index numbers created from the *Private Non-Residential Investment Deflator*. Prices of domestic corporate goods and major goods included in the *Machinery Statistics* were compared and brought into alignment to create the indexes. Weights were derived from the production value listed in the *Machinery Statistics* (Fig. 2.14).

2.14.3 Fisher's Ideal Formula

Sir, Ronald Aylmer Fisher, is the Father of Statistics for his immense contribution in the field of applying and discovering statistical methods in genetics and the design of experiments.

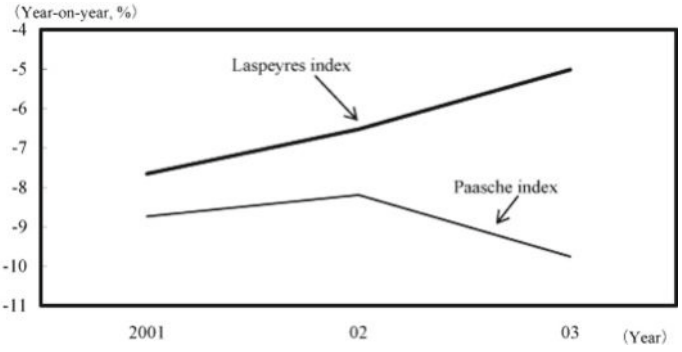


Fig. 2.14 Visual representation of Laspeyre's and Paasche's index numbers

Hald, Anders mentioned Fisher as "A genius who almost single-handedly created the foundations for modern statistical science" in their book, *A History of Mathematical Statistics (1998).*

He also formulated a weighted aggregate index number using a geometrical average of Paasche's and Laspeyre's indices (Fig. 2.15).

Fisher's Ideal Index

$$= \sqrt{\text{Paasche's Index} * \text{Laspeyr's Index}} * 100$$

$$= \sqrt{\frac{\sum(X_{c,t_n}) * \sum(X'_{c,t0})}{\sum(X_{c,t_0}) * \sum(X'_{c,t_n})} * \frac{\sum(X_{c,t_n}) * \sum(X'_{c,t_n})}{\sum(X_{c,t_0}) * \sum(X'_{c,t_n})}} * 100$$

where X' is the weights for calculating index number of X. (c, t_n) refers to cost at time t_n and (c, t_0) refers to cost in the base year.

Therefore, the price index number defined as per Fisher's formula is given as,

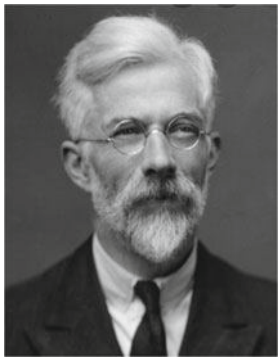


Fig. 2.15 Sir Ronald A Fisher

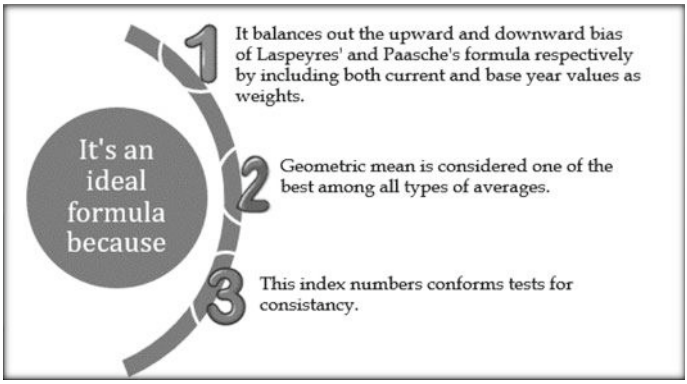


Fig. 2.16 Why is Fisher's Index the ideal one

$$P_{01}^{F} = \sqrt{\frac{\sum p_1 q_1}{\sum p_0 q_1} * \frac{\sum p_1 q_0}{\sum p_0 q_0}} * 100$$

And the quantity index number would be, $Q_{01} F \frac{\sum q_1 p_1}{\sum q_0 p_1} * \frac{\sum q_1 p_0}{\sum q_0 p_0}$.

Note: Why is Fisher's index called an ideal one?

This method of calculation reveals the unbiased value of today's basket of goods being constructed in constant dollars. This formula evens out the biasedness of understating and overstating inflation by reducing it into half by choosing a geometric mean. This method also allows for some variations in the characteristics of the base weights (Fig. 2.16).

Demerits

Boddington has remarked that "Unfortunately, while this formula apparently meets most of the mathematical requirements of a perfect index formula, it is objected to on the score that it is not clear what it measures, i.e., the result combines both price and volume changes when usually we want the one to be separated from the other". Few statisticians are of an opinion that there exists a need for an alternative and all-purpose formula for calculating weighted aggregate index numbers. This is because Fisher's method of calculation goes pointless even with one missing value. We must also note that the collection of all the data points may be time consuming and expensive at times. Though the calculation is not tedious, ambiguity arises when both current and base year values are used as weights as we cannot confidently identify the cause of inflation.

2.14.4 Dorbish and Bowley

These were economists who considered using the arithmetic average of both Laspeyre's and Paasche's indices.

$$\frac{\text{Paasche's Index} + \text{Laspeyr's Index}}{2} = \frac{\frac{\sum(X_{c,t_n}) * \sum(X'_{c,t_0})}{\sum(X_{c,t_0}) * \sum(X'_{c,t_0})} + \frac{\sum(X_{c,t_n}) * \sum(X'_{c,t_n})}{\sum(X_{c,t_0}) * \sum(X'_{c,t_n})}}{2}$$

where X' is the weights for calculating the index number of X. (c, t_n) refers to cost at time t_n and (c, t_0) refers to cost in the base year. For example, the price index number defined by Dorbish and Bowley would be,

$$P_{01}^{\text{DB}} = \left(\frac{P_{01}^{L} + P_{01}^{P}}{2}\right) * 100 = \frac{\frac{\sum p_1 q_0}{\sum p_0 q_0} + \frac{\sum p_1 q_1}{\sum p_0 q_1}}{2} * 100$$

But the result was vulnerable to the outliers and the simple average was pointless in many cases.

2.14.5 Marshall–Edgeworth's Index

To overcome the drawbacks of Dorbish and Bowley, Economic Statisticians, Marshall (1887) and Edgeworth (1925) considered the total of both current and base year quantities.

$$\text{Marshall-Edgeworth's Index} = \frac{\sum X_{c,t_n} * \left(X'_{c,t_0} + X'_{c,t_n}\right)}{\sum X_{c,t_0} * \left(X'_{c,t_0} + X'_{c,t_n}\right)}$$

where X' is the weights for calculating the index number of X. (c, t_n) refers to cost at time t_n and (c, t_0) refers to cost in the base year. For example, the price index number is defined as,

$$P_{01}^{\text{ME}} = \left[\frac{\sum p_1(q_0 + q_1)}{\sum p_0(q_0 + q_1)}\right] * 100 = \left[\frac{\sum p_1 q_0 + p_1 q_1}{\sum p_0 q_0 + p_0 q_1}\right] * 100$$

2.14.6 Kelly's Index Number

Truman L Kelly suggested considering a weight that is an arithmetic average of base and current year quantities. Consider $\overline{X'}$ as an arithmetic mean of weights $\left(X'_{c,t_0} + X'_{c,t_n}\right)/2$, the formula is thus, $\frac{\sum X_{c,t_n} * \overline{X'}}{\sum X_{c,t_0} * \overline{X'}}$.

where X' is the weights for calculating the index number of X. (c, t_n) refers to cost at time t_n and (c, t_0) refers to cost in the base year. For example, the price index number is defined as,

$$P_{01}^{K} = \frac{\sum p_1 * \overline{q}}{\sum p_0 \overline{q}} * 100 \quad \text{where } \overline{q} = \frac{q_1 + q_0}{2}$$

Though this idea overcomes a few disadvantages of Dorbish's and Bowley's indices, yet the application remained inappropriate in many real-time scenarios. Therefore, economists experimented with geometric and harmonic mean of weights.

2.14.7 Walsh's Index Number

Walsh was one such economist who modified Kelly's index by using a geometric average instead of an arithmetic mean. Geometric mean in general is more accurate than simple average since it considers the compounding effect of two time periods. Another important advantage of this formula is the base period which can be changed without necessitating a corresponding change in the weights.

Consider $\tilde{X}' = \sqrt{X'_{c,t_0} * X'_{c,t_n}}$, the formula is thus, $\frac{\sum X_{c,t_n} * \tilde{X}'}{\sum X_{c,t_0} * \tilde{X}'}$.

where X' is the weights for calculating the index number of X. (c, t_n) refers to cost at time t_n and (c, t_0) refers to cost in the base year. For example, the price index number is defined as,

$$P_{01}^{W} = \frac{\sum p_1 * \widetilde{q}}{\sum p_0 * \widetilde{q}} * 100 \quad \text{where } \widetilde{q} = \sqrt{q_1 * q_0}.$$

2.15 Criteria for a Good Index Number

Despite all our efforts, there can be errors in the construction of index numbers. These are mainly classified into

– Homogeneity errors: In such cases, the composition of commodities of index numbers is not on the same level.

– Sampling errors: In certain situations where a sample of prices/quantities is collected to summarize the price/quantity index numbers, there are chances of the sample not being representative of the population. This gives rise to sampling errors.
– Formula errors: Formula errors can be such as the application of the incorrect formula.

2.15.1 Tests on Index Numbers

These are statistical techniques used to assess the quality and reliability of index numbers, which are used to measure changes in a particular variable over time. Index numbers are essential for comparing economic, financial, or other data across different periods. Some common tests applied to index numbers include:

1. Time-Reversal Test: The Time-Reversal Test is used to check the reliability and consistency of an index by reversing the time-series data and recomputing the index. If the index is accurate, the value obtained from the reversed data should be approximately the reciprocal of the original index. In mathematical terms, if the original index is $I(t)$, then the time-reversed index would be $1/I(t)$.
2. Factor Reversal Test: The factor reversal test is similar to the time-reversal test, but it involves reversing the weights or factors used to construct the index. By reversing the weights and recomputing the index, you can check if the index remains consistent and reliable.
3. Resampling Test: The resampling test involves randomly reshuffling the data points to create new datasets. Then, you recalculate the index using each of these new datasets and compare the results to the original index. This test helps assess the stability and robustness of the index.
4. Linking Test: The linking test is conducted when there is a change in the base period of the index. It involves comparing the index values before and after the change to ensure the continuity and consistency of the index series.
5. Time-Series Analysis: Various time-series analysis techniques, such as autocorrelation and seasonality analysis, can be applied to index numbers to identify patterns, trends, and potential data issues.
6. Unit Test: The unit test is performed to ensure that the index remains unaffected when the unit of measurement is changed. For example, if the index is calculated in dollars, the unit test checks if the index remains the same when converted into another currency.
7. Plausibility Check: Plausibility checks involve comparing the index results with other economic indicators and related data to see if they align logically and are reasonable.

It is important to note that these tests are not exhaustive, and the choice of specific tests may vary depending on the nature of the data and the purpose of the index. The ultimate goal is to ensure the accuracy, consistency, and reliability of the index

numbers so that they can be used effectively for various economic and analytical purposes.

2.15.2 Time-Reversal Test (T.R.T.)

Does your index number work both the ways, forward and backward?

I_{01} =Is the index number calculated for time "1" with base time "0." Similarly, I_{10} is the index number for year "0" on year "1."

The time-reversal test is thus expressed as, $I_{01} * I_{10} = 1$.

The index number calculated backward is the reciprocal of the index number calculated forward. Fisher's index number satisfies the Time-Reversal Test.

The analogy goes this way, and the price of a commodity ranges from Rs 5 to Rs 20 between the years, 1980 and 1990. This means that in the year 1990, the product is 4 times that of the price in 1980. In the year 1980, the product price is only 25% of the cost in 1990.

The product $4 * 0.25 = 1$.

Let us decipher through the formula, consider the price index number given by Fisher,

$$P_{01}^{P} = \sqrt{\frac{\sum p_1 q_1}{\sum p_0 q_1} * \frac{\sum p_1 q_0}{\sum p_0 q_0}}$$

When time subscripts are interchanged, we get,

$$P_{10}^{P} = \sqrt{\frac{\sum p_0 q_0}{\sum p_1 q_0} * \frac{\sum p_0 q_1}{\sum p_1 q_1}}$$

$$P_{01}^{P} * P_{10}^{P} = 1$$

Therefore, the Time-Reversal Test is satisfied.

The other methods that satisfy the test are,

- The simple geometric mean of price relatives.
- Marshall–Edgeworth formula of index numbers.

– Kelly's method as it is a weighted geometric mean of price relatives with fixed weights.
– Fisher's index numbers.

2.15.3 Factor Reversal Test (F.R.T.)

Does your index number validate the change in the value according to the change in price and quantity?

The idea behind the test has a very simple logic. Let us consider, P_{01} and Q_{01} are the price and quantity index numbers that are defined at period "1" with "0" as the base period, respectively. V_{01} is defined as a value index number which is defined as $V_{01} = P_{01} * Q_{01}$. Let us take an example from history where machinery revolutionized the handcraft industry. During the period between 1990 and 1995, the cost of hand-weaved baskets increased by 20% and the production surged by 9 times. This means the value index number must indicate an 18% rise (20% rise in price and 9 times rise in production levels) to satisfy the factor reversal test.

Let us look at why Laspeyre's index doesn't satisfy the F.R.T., but only Fisher's method of index numbers satisfies F.R.T. Let's consider Laspeyre's method.

$$P_{01}^{L} = \frac{\sum p_1 q_0}{\sum p_0 q_0}, Q_{01}^{L} = \frac{\sum q_1 p_0}{\sum q_0 p_0} V_{01}^{L} = \frac{\sum q_1 p_1}{\sum q_0 p_0}$$

When we multiply $P_{01}^{L} * Q_{01}^{L} = \frac{\sum p_1 q_0}{\sum p_0 q_0} * \frac{\sum q_1 p_0}{\sum q_0 p_0} = \frac{\sum p_1 q_0 * \sum q_1 p_0}{\left(\sum p_0 q_0\right)^2} \neq V_{01}^{L}$. On the other hand, let's look at Fisher's method,

$$\sqrt{\frac{\sum (X_{c,t_n}) * \sum \left(X'_{c,t_0}\right)}{\sum (X_{c,t_0}) * \sum \left(X'_{c,t_0}\right)} * \frac{\sum (X_{c,t_n}) * \sum \left(X'_{c,t_n}\right)}{\sum (X_{c,t_0}) * \sum \left(X'_{c,t_n}\right)}}$$

$$P_{01}^{F} * Q_{01}^{F} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} * \frac{\sum q_1 p_1}{\sum q_0 p_1} * \frac{\sum p_1 q_1}{\sum p_0 q_1} * \frac{\sum p_1 q_0}{\sum p_0 q_0}}$$

$$= \sqrt{\frac{\sum p_1 q_1}{\sum p_0 q_0} * \frac{\sum q_1 p_1}{\sum q_0 p_0}} = \frac{\sum q_1 p_1}{\sum q_0 p_0}$$

which is the value index defined.

Let us compute the Fisher's index number for data collected from a household on their grocery with specific units of measurement and understand how the formula satisfies T.R.T. and F.R.T. (Tables 2.15 and 2.16).

Solution

The Fisher's index number is calculated as,

For Time-Reversal Test,

$$P_{01}^{F} = \sqrt{\frac{\sum p_1 q_1}{\sum p_0 q_1} * \frac{\sum p_1 q_0}{\sum p_0 q_0}} * 100 \quad \text{and} \quad P_{01}^{F} = \sqrt{\frac{\sum p_0 q_0}{\sum p_1 q_0} * \frac{\sum p_0 q_1}{\sum p_1 q_1}} * 100$$

$$\text{T.R.T.} = P_{01}^{F} * P_{01}^{F} = 1 \text{ (ignoring the 100)}$$

$$\text{T.R.T.} = \left\{\sqrt{\frac{8400}{21280} * \frac{22505.5}{16756}}\right\} * \left\{\sqrt{\frac{16756}{22505.5} * \frac{16756}{8400}}\right\} = \sqrt{1} = 1$$

For factor reversal test,

Table 2.15 Price and quantity of products along with base and current years

Items	Base year		Current year	
	Quantity	Price	Quantity	Price
Wheat	12	326	15	560
Rice	15	550	18	650
Oil	5	250	9	354
Pulses	8	400	10	512
Spices	1.5	96	2.5	113

Table 2.16 Calculation for time and factor reversal tests

Items	Base year		Current year					
	Quantity q_0	Price p_0	Quantity q_1	Price p_1	p_1q_0	p_0q_0	p_1q_1	p_0q_1
Wheat	12	326	15	560	6720	3912	8400	4890
Rice	15	550	18	650	9750	8250	11,700	9900
Oil	5	250	9	354	1770	1250	3186	2250
Pulses	8	400	10	512	4096	3200	5120	4000
Spices	1.5	96	2.5	113	169.5	144	282.5	240
Total	41.5	1622	54.5	2189	22,505.5	16,756	28,688.5	21,280

$$\text{F.R.T.} = V_{01}^{F} = P_{01}^{F} * Q_{01}^{F} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$$

$$P_{01}^{F} = \sqrt{\frac{\sum p_1 q_1}{\sum p_0 q_1} * \frac{\sum p_1 q_0}{\sum p_0 q_0}} \text{ and } Q_{01}^{F} \sqrt{\frac{\sum q_1 p_1}{\sum q_0 p_1} * \frac{\sum q_1 p_0}{\sum q_0 p_0}}$$

$$P_{01}^{F} * Q_{01}^{F} = \left\{ \sqrt{\frac{28688.5}{21280} * \frac{22505.5}{16756}} \right\} * \left\{ \sqrt{\frac{28688.5}{22505.5} * \frac{21280}{16756}} \right\}$$

$$= \left\{ \sqrt{\left(\frac{28688.5}{16756} \right)^2} \right\} = \frac{28688.5}{16756} = \frac{\sum p_1 q_1}{\sum p_0 q_{10}} = V_{01}^{F}$$

Thus, Fisher's index satisfies both T.R.T and F.R.T.

2.15.4 Unit Test

This test states that the method of constructing index numbers should be independent of units of measurement. This makes comparison simpler by bringing all the values to equal levels. For example, consider the entities in any household budget, petrol is in liters, rice is in quintals, electricity in watts consumed, cloth in meters, and others. Since the rule is very simple, all the methods except for the simple aggregate method will satisfy the unit test.

2.15.5 Circular Test

The circular test is an extension of the Time-Reversal Test (TRT). Consider a period of 2 years with three index values computed based on the shiftability of the base period. For example, if an index is constructed for the year 2015 with the base of 2014 and another index for 2014 with the base of 2013, then it should be possible for us to directly get an index for the year 2015 with the base of 2013. If the index calculated directly does give a consistent value, then the circular test is said to be satisfied.

$$P_{01} * P_{12} * P_{20} = 1$$

The main idea behind this test is to check if the index number is adhered to the changes in time even base period is not fixed.

- Simple aggregative method.
- Fixed weight aggregative method.

– Kelly's method.

In a simple aggregative method,

$$P_{01} = \frac{P_1}{P_0}, P_{12} = \frac{P_2}{P_1}, P_{20} = \frac{P_0}{P_2}$$

$$P_{01} * P_{12} * P_{20} = \frac{P_1}{P_0} * \frac{P_2}{P_1} * \frac{P_0}{P_2} = 1$$

In the fixed weight aggregative method, which also satisfies circular test, weights q will be added to each price element.

2.16 Shifting the Base Year

Practically, index number calculations can be made more meaningful when base years are dynamic than fixed. This is because with the passage of time the base values are also affected due to numerous economic and social factors (Fig. 2.17).

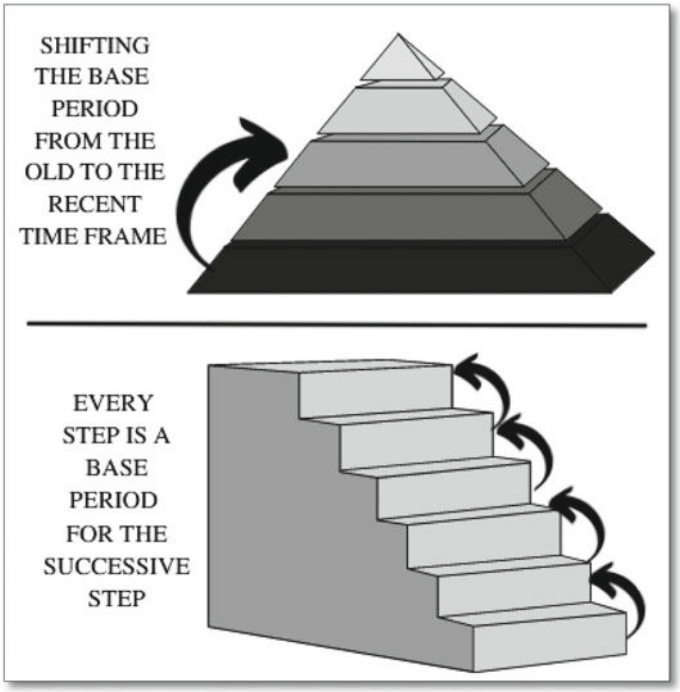


Fig. 2.17 Visual explanation of base shifting

Table 2.17 Index numbers for the year 2010 till 2014

Year	2010	2011	2012	2013	2014
Index	100	120	124	136	140

Table 2.18 Index number calculation with base period 2010 and 2012

Year	Base 2010	Base 2012
2010	100	(100/100) * 100 =80.64
2011	120	(120/100) * 100 =120
2012	124	(124/120) * 100 =103.33
2013	136	(136/103.33) * 100 =131.61
2014	140	(140/131.61) * 100 =106.37

There can be two scenarios in which we can alter the base period (Tables 2.17 and 2.18).

– Using a fixed base period approach, we move from an older to a newer base. For example, few socioeconomic indexes such as Cost-of-Living Index, Health Index, Labor Index, and Happiness Index are calculated with a difference of 10-year base period starting from the year of the census enumeration.

2.17 Chain Base Index and Link Relatives

Chain base index numbers are a method used in economics and statistics to calculate the relative changes in a particular variable's value over time. The process involves setting a particular period as the base period and then comparing the values of subsequent periods with the base period. The base period value is given a base index number of 100 to facilitate comparisons.

For example, if you have data on the prices of a basket of goods for different years, you can choose one year (e.g., 2020) as the base year. Then, you calculate the price index for each subsequent year by dividing the price of the basket of goods in that year by the price of the same basket in the base year (2020) and multiplying by 100.

Chain base index numbers are particularly useful when comparing changes over time and avoiding the problem of a fixed base period, which can become outdated and not reflective of the current economic situation.

Link relatives are a way to express the percentage change in a variable between two adjacent periods, usually months or years. They are calculated by taking the ratio of the value in the current period to the value in the previous period and then expressing the change as a percentage.

The formula to calculate the link relative is:

Table 2.19 Index number calculation using chain base index numbers

Year	2010	2011	2012	2013	2014
Index	100	120	124	136	140
Chain-based index numbers	100	(100/120) * 100 = 83.33	(120/124) * 100 = 96.77	(124/136) * 100 = 91.17	(136/140) * 100 = 97.14

Table 2.20 Calculation of chain indices for Channapatna toys

Year	2000	2001	2002	2003	2004	2005
Chain indices	108	85	67	96	102	88

$$\text{Link Relative} = (\text{Value in the current period}/\text{Value in the previous period}) * 100$$

Link relatives are useful for understanding short-term changes in a variable, but they don't take into account long-term trends or the cumulative effect of multiple changes over time.

Chain base index numbers and link relatives are both methods used to analyze changes in variables over time. Chain base index numbers are useful for understanding the relative changes in a variable with respect to a chosen base period, while link relatives provide a quick way to analyze the percentage change between two adjacent periods.

Example: For the following data let us see how to shift the base from 2010 to 2012.

Chain Base Index Number:

Example: At the famous Channapatna toys, we have collected the chain indices of wooden toys for Montessori kids. From the chain base index numbers given below, obtain the fixed base index numbers to help a researcher understand the purchasing pattern of the toys (Tables 2.19 and 2.20).

Solution

(Tables 2.21 and 2.22).

Table 2.21 Calculation of fixed based index numbers from chain indices

Year	Chain index number	Fixed base index number
2000	108	108
2001	85	(108 ×85/100) =**92**
2002	67	(92 ×67/100) =**62**
2003	96	(62 ×96/100) =**60**
2004	102	(60 ×102/100) =**61**
2005	88	(61 ×88/100) =**54**

Table 2.22 Calculation of link relatives and chain relatives for prices of jute and tea from 2002 to 2005

Particulars		Year			
		2002	2003	2004	2005
Jute	Quantity (in 100 kg.)	851	725	712	612
	Price (Rs.100 kg.)	215	325	304	296
Tea	Quantity (in 100 kg.)	188	172	230	196
	Price (Rs/per 100 kg.)	556	755	896	802

Solution

(Tables 2.23 and 2.24).

Example 2: The database given below is the wholesale prices from a petrochemincal manufacturing industry. Construct chain index numbers for the year 2008=09 to 2011–12 from the following Table 2.24.

Solution

Link Relatives

$$\text{Link relatives for year } 2009-2010 = \frac{153.8}{174.6} * 100 = \mathbf{113.52}$$

$$\text{Link relatives for year } 2009-2010 = \frac{144.5}{168.2} * 100 = \mathbf{116.40}$$

$$\text{Link relatives for year } 2009-2010 = \frac{166.6}{172.5} * 100 = \mathbf{103.54}$$

$$\text{Link relatives for year } 2009-2010 = \frac{162.8}{176.2} * 100 = \mathbf{108.23}$$

Chain Relatives (Table 2.25)

$$\text{Chain index for year } 2009-2010 = \frac{100}{110.42} * 100 = \mathbf{110.42}$$

$$\text{Chain index for year } 2010-2011 = \frac{110.42}{118.35} * 100 = \mathbf{130.68}$$

$$\text{Chain index for year } 2011-2012 = \frac{130.68}{110.72} * 100 = \mathbf{144.69}$$

Table 2.23 Calculation of link relatives and chain relatives for prices of jute and tea from 2002 to 2005

Jute (base 2002)	2002	2003	2004	2005
Quantity (q)	815	725	712	612
Price (p)	215	325	304	296
Value ($q \times p$)	182,965	235,625	216,448	181,152
Index number(I_1)	100	128.78	118.30	99.01
Tea (base 2002)				
Quantity (q)	188	172	230	196
Price (p)	556	755	896	802
Value ($q \times p$)	104,528	129,860	206,080	157,192
Index number(I_2)	100	124.23	197.15	150.38
Sum of Index Numbers $= (I_1 + I_2)$	200	253.02	315.45	249.39
Average of Index Numbers $= (I_1 + I_2)/2$	100	126.51	157.73	124.70
Link relatives	100	$(126.51/100)^*100$ $=$**126.51**	$(157.73/126.51) * 100$ $=$**124.68**	$(124.7/157.73)$ $*100 = 79.06$
Chain relatives	100	$\big(100^*126.51\big)/100$ $=$**126.51**	$126.51 * 124.68)/100 =$**157.73**	$(79.06 * 157.73)/100 =$**124.70**

Table 2.24 Indices for wholesale prices of petrochemical manufacturing industry

Year	Index numbers of wholesale prices			
	Primary articles (A)	Fuel group (B)	Manufactured products (C)	All commodities (D)
2008–2009	153.8	144.5	166.6	162.8
2009–2010	174.6	168.2	172.5	176.2
2010–2011	208.2	196	203.8	210.5
2011–2012	224.4	228.2	222.2	230.8

Table 2.25 Calculation of chain indices for wholesale prices of petrochemical manufacturing industry

	Link relatives			
	2008–2009 (Base Year)	2009–2010	2010–2011	2011–2012
A	100	113.52	119.24	107.78
B	100	116.40	116.53	116.43
C	100	103.54	118.14	109.03
D	100	108.23	119.47	109.64
Total of L.R.	**400**	**441.70**	**473.38**	**442.88**
Average of L.R.	**100**	**110.42**	**118.35**	**110.72**
Chain index	**100**	**110.42**	**130.68**	**144.69**

2.18 Splicing of Index Numbers

Splicing refers to a method in which two datasets are combined by merging the base values. When comparing two series of datasets, it is always recommended to have the same base periods since this puts the datasets at equal levels. Take a look at how the splicing of data is done. (Table 2.26).

Example: Consider the old and new series in the following problem. The solutions are explained in the table with calculations (Table 2.27).

Example: Consider this information where the aggregate values of prices and quantities are mentioned. Let us simplify the data by calculating two price indices in Series A with q_o as weights and Series B with q_3 as weights. Using the data let us combine both the series into one continuous series by splicing.

Table 2.26 Method of splicing data

Splicing method	Old series	New Series
Forward	$\left(\frac{100}{\text{Index number of old series}}\right)$*Given index of old series	No change
Backward	No change	$\left(\frac{\text{Index number of old series}}{100}\right)$*Given index of new series

Table 2.27 Example of splicing data with Series A and Series B

Year	Series A	Series B	Series B spliced to series A	Series A spliced to series B
1984	100		100	$\frac{100}{160} \times 100 = \mathbf{62.5}$
1985	120		120	$\frac{100}{160} \times 120 = \mathbf{75}$
1986	140		140	$\frac{100}{160} \times 140 = \mathbf{87.5}$
1987	160	100	$\frac{160}{100} \times 100 = 160$	$\frac{100}{160} \times 160 = \mathbf{0}$
1988		115.56	$\frac{160}{100} \times 115.56 = \mathbf{184.89}$	115.56
1989		128.89	$\frac{160}{100} \times 128.89 = \mathbf{206.22}$	128.89
1990		144.44	$\frac{160}{100} \times 144.44 = \mathbf{231.11}$	144.44

2.19 Deflating Index Numbers

Deflating index numbers are a statistical technique used to adjust or remove the effect of price changes from a specific economic indicator or index. It allows for a more accurate comparison of the underlying data over time by eliminating the impact of inflation or deflation.

Index numbers are often used to measure changes in economic variables like consumer prices, production levels, or economic growth. However, over time, these index numbers may be affected by changes in the general price level, making it difficult to interpret the true underlying trends. To address this, deflating index numbers involve dividing the index number by a price index, typically a measure of the average price level over a given period. The resulting deflated index provides a value that represents the indicator's real changes, adjusting for the impact of price fluctuations.

The formula for deflating an index number is as follows:

$$\textbf{Deflated Index} = (\textbf{Nominal Index}/\textbf{Price Index}) * \mathbf{100}$$

Here's a step-by-step explanation of how to deflate an index number:

1. Obtain the nominal index: This is the original index you want to adjust for inflation or deflation. For example, it could be a Consumer Price Index, a GDP index, or a production index.
2. Obtain the price index: This is the average price level for the corresponding period, often represented as a price index like the Consumer Price Index (CPI) or Producer Price Index (PPI).
3. Divide the nominal index by the price index: This step involves dividing the nominal index number by the price index value for the same time period.
4. Multiply by 100: This step is optional but is often done to express the deflated index as a percentage.

The resulting deflated index provides a more accurate representation of the changes in the underlying indicator, eliminating the influence of price changes. This

allows for better comparisons of data over time and provides a clearer picture of actual economic trends.

Deflating index numbers are commonly used in economic analysis, financial reporting, and other fields where accurate comparisons across time periods are necessary (Table 2.28).

Example: Consider the physical volume of inventory at paint manufacturing unit in Rajasthan. The data for 6 years is given below, and with respect to base year 2010, compute the deflated inventory.

Solution

Deflated income is calculated as: (Inventory/WPI) * 100.

Consider 2005 year as base period. The volume of inventory will be 100. As the years progressed, the levels decreased (Table 2.29).

Volume of inventory is calculated as:

$\frac{100}{\text{deflated inventory of base year}}$*Deflated inventory of the current year .

For example, volume of inventory in year 2007 is calculated as, (100/726.06) * 525.82 =72.42.

2.20 Note on Real Income

Real income refers to the purchasing power of the income. In simple terms, inflation-adjusted income is called real income.

Real income

$$\textbf{\textit{Real income}} = \frac{\textbf{\textit{Nominal income}}}{\textbf{\textit{CPI}}}$$

Example: Consider a scenario where the CPI and nominal income of an employee are inversely proportional.

As we can notice from Table 2.30, though the earnings have certainly increased, the real income of the person has decreased because the purchasing power of the money is reduced. (Real income < nominal income for the year 2000 and 2005.) This means that the goods worth 36,585.37 in the year are reduced to 22,500.

Example: The annual wages (in Rs.) of workers are given along with consumer price indices. Find: (i) the real wages, (ii) the real wage indices (Tables 2.31 and 2.32).

Solution

Additional Notes

In summary, index numbers contribute significantly to nation building by providing essential information for economic monitoring, policy formulation, resource allocation, business decision-making, investor confidence, and tracking progress toward

Table 2.28 Inventory and WPI of paints in a manufacturing unit in Rajasthan

Year	2005	2006	2007	2008	2009	2010
Inventory (thousands)	785.6	647.8	578.4	464.6	441.7	400.8
WPI	108.2	105.5	110	103.9	110.6	111.5

Table 2.29 Physical volume of inventory at paint manufacturing unit in Rajasthan

Year	Inventory (thousands)	W.P.I.	Deflated inventory	Volume of inventory
2005	785.6	108.2	726.06	100.00
2006	647.8	105.5	614.03	84.57
2007	578.4	110	525.82	72.42
2008	464.6	103.9	447.16	61.59
2009	441.7	110.6	399.37	55
2010	400.8	111.5	359.46	49.51

Table 2.30 Real income

Year	CPI (approx.)	Nominal income	Calculation	Value
1980	82	Rs. 30,000	(30,000/82) * 100	Rs. 36,585.37
2000	172	Rs. 40,000	(40,000/172) * 100	Rs. 23,255.81
2005	200	Rs. 45,000	(45,000/200) * 100	Rs. 22,500

Table 2.31 Wages and CPI of workers

Year	2001	2002	2003	2004
Wages	1890	1756	2345	3214
CPI	100	145	350	150

Table 2.32 Calculation of real wages using annual wages and CPI of workers

Year	Wages	Price index	Real wages =(Wages/Price Index) * 100	Real wage 1980 =100
2001	1890	100	(1890/100) * 100 =1890	(1890/1890) * 100 =**100**
2002	1756	145	(1756/145) * 100 =1211.03	(1211.03/1890) * 100 =**64.07**
2003	2345	350	(2345/350) * 100 =670	(670/1890) * 100 =**35.44**
2004	3214	150	(3214/150) * 100 =2142.67	(2142.67/1890) * 100 =**113.36**

socioeconomic development goals. By leveraging accurate and reliable index numbers, nations can make informed decisions, promote inclusive growth, and work toward sustainable development and prosperity.

Some real-life examples of how index numbers have been used for policymaking:

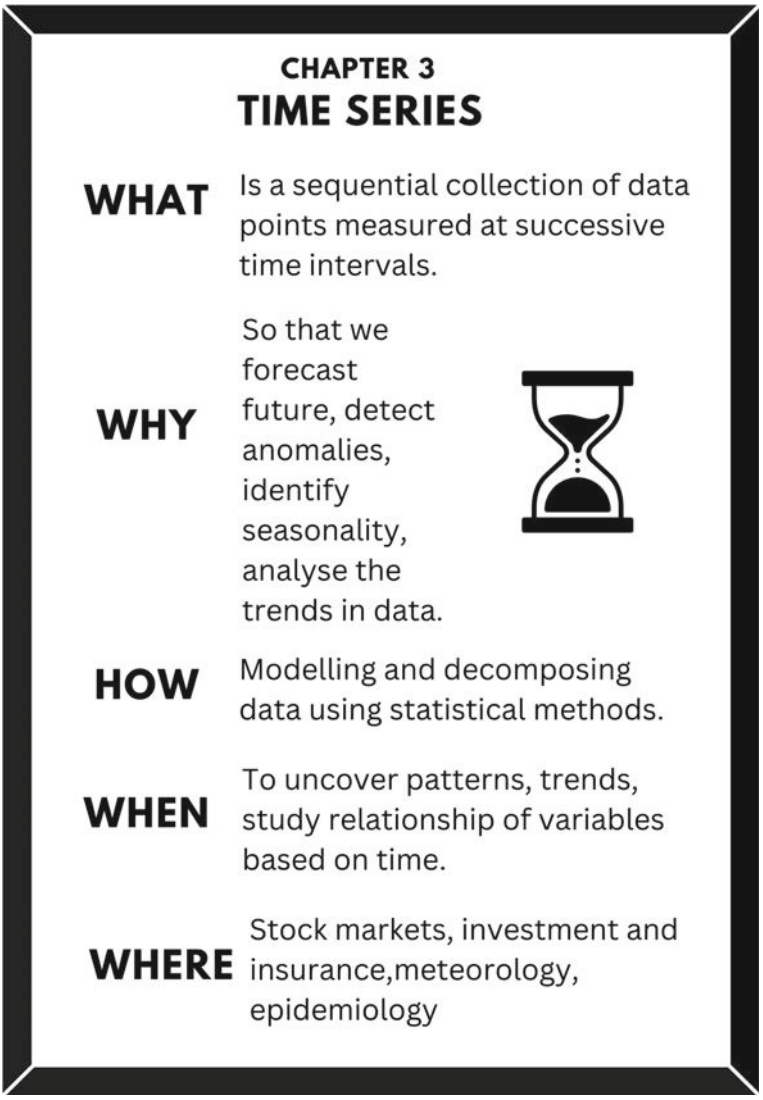
1. Using Consumer Price Index (CPI) for Monetary Policy: Central banks often use the CPI to gauge inflationary pressures and adjust monetary policy accordingly. For instance, the Reserve Bank of India (RBI) relies on CPI numbers to set interest rates. In 2016, the RBI adopted a policy framework that targeted a specific inflation rate based on CPI. By closely monitoring the CPI, the central bank

can make decisions on whether to tighten or loosen monetary policy to control inflation and promote price stability.

2. Adjusting Social Security Benefits using Index Numbers: Governments use index numbers to adjust social security benefits and pensions to account for changes in the cost of living. For example, in the United States, social security benefits are annually adjusted based on the changes in the CPI. This ensures that the purchasing power of retirees and social security recipients is maintained as the cost of living changes over time.
3. Infrastructure Investment based on Infrastructure Index: Countries often develop infrastructure indices to assess the quality and performance of their infrastructure systems. These indices help policymakers identify areas requiring investment and prioritize infrastructure development projects. For instance, the Global Infrastructure Hub developed the Global Infrastructure Outlook, which uses an index to assess the infrastructure needs of various countries and guide investment decisions.
4. Using Education Index to Monitor Education Policy: Education indices, such as the Programme for International Student Assessment (PISA) index, are used to assess the quality of education systems across countries. Policymakers can utilize the insights from these indices to identify areas for improvement, formulate education policies, and allocate resources effectively. For example, countries like Finland and South Korea have used PISA scores to inform education reforms and improve educational outcomes.
5. Monitoring Sustainable Development Goals (SDGs): Index numbers are instrumental in tracking progress toward the SDGs. For instance, the United Nations Development Programme (UNDP) publishes the Human Development Index (HDI) to measure and compare countries' progress in areas such as health, education, and income. Policymakers use HDI data to assess the impact of policies on human development, identify areas for improvement, and allocate resources accordingly.

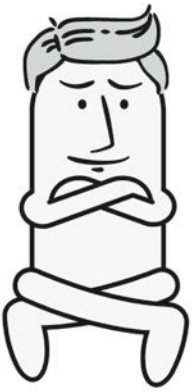
These examples demonstrate how index numbers are utilized to inform policy decisions in various domains, such as monetary policy, social security, infrastructure investment, education, and sustainable development. By leveraging index numbers, policymakers can make data-driven decisions, target interventions effectively, and monitor the outcomes of their policies.

Chapter 3
Time Series



S. Prasad, *Advanced Statistical Methods*, https://doi.org/10.1007/978-981-99-7257-9_3

TIME SERIES



Mr STAT

- Introduction.
- Mathematical models of time series.
- Components of time series.
- Decomposition of time series.
- Stochastic process.
- Autocorrelation
- ARIMA Models.

Miss TICS

- Decomposition of time series with 'international airline passengers' time series data set for the year 1949 to 1960.
- Learn about Moving Average Convergence Divergence (M.A.C.D graph)
- Help a farmer check if excessive use of fertilisers has decreased the rice production

What is a time series?

A time series is a collection of data points ordered chronologically and recorded at successive time intervals. These data points can be taken over various frequencies, such as seconds, minutes, hours, days, months, or years. Time series data is commonly used in fields like economics, finance, weather forecasting, stock market analysis, epidemiology, and more.

Why is time series analysis important?

Time series analysis provides valuable insights into the underlying patterns and trends present in the data. By understanding these patterns, businesses and researchers can make informed decisions, forecast future values, detect anomalies, identify seasonality, and perform trend analysis. It helps in understanding the dynamic behavior of a system and can be used for predictive modeling.

How is time series analysis conducted?

Time series analysis involves several steps, including:

- Data collection: Gathering data at regular intervals.
- Data cleaning: Handling missing values, outliers, and errors.
- Visualization: Plotting the data to observe trends and patterns visually.
- Decomposition: Separating the data into trend, seasonality, and residual components.
- Modeling: Applying statistical methods like ARIMA (AutoRegressive Integrated Moving Average), SARIMA (Seasonal ARIMA), or machine learning algorithms for forecasting.
- Evaluation: Assessing the model's performance and making adjustments if necessary.

When to use time series analysis?

Time series analysis is used when the data points have a temporal ordering and the focus is on understanding how the data changes over time. It is employed in scenarios where past values are indicative of future behavior, such as stock market predictions, weather forecasting, sales forecasting, and more. Time series analysis is not suitable for non-sequential data, where the order of observations does not matter.

Where is time series analysis applied?

Time series analysis finds applications in various fields, including:

- Finance: Analyzing stock prices, currency exchange rates, and economic indicators.
- Economics: Studying GDP, inflation rates, and unemployment data.
- Meteorology: Forecasting weather patterns and climate change trends.
- Epidemiology: Tracking disease outbreaks and analyzing health data.
- Operations Research: Optimizing inventory management and supply chain forecasting.

In this dynamic world, we notice that many aspects of life change over time. Time series analysis is a statistical technique used to analyze and understand patterns in sequential data that is collected over time. Time series analysis is a complex field that requires a deep understanding of statistical modeling and data analysis techniques.

3.1 Definition

A time series consists of statistical data which are collected, recorded, or observed over successive increments.—Patterson.

A time series is a set of statistical observations arranged in chronological order.—**Morris Hamburg.**

Time series refers to the collection of observations through repeated measurements over time. There are two ways in which information can be collected based on the interval of time which is REGULAR and IRREGULAR. As shown in Fig. 3.1 depicts time series data collected at regular and irregular intervals of time.

Before we start with any statistical methods, we must always ensure that the data collected is based on a specific time period, like hours/seconds/month/year/bi-annual/quarterly, etc. In general, time is plotted on the X axis of the graph.

Examples:

- One best example of time series charts is the one collected from an ECG device, (Echocardiogram) which constantly monitors the activity of the heart.
- Sensex is also an example of how stock prices fluctuate during the hours of trade.
- Another familiar time series graph is the functioning of CPU and cache memory shown by the control panel.

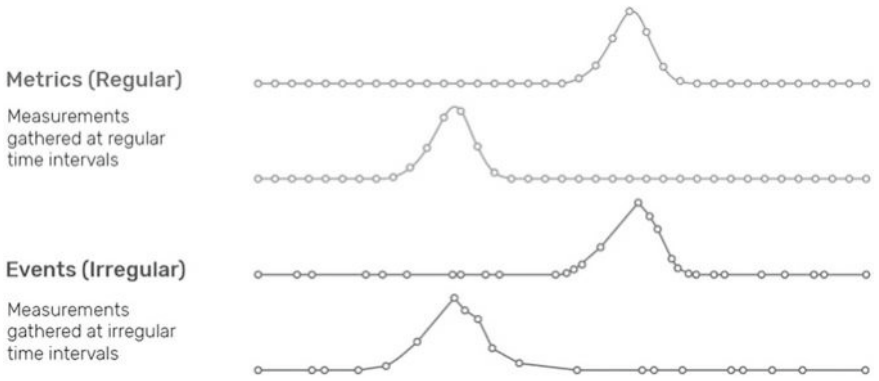


Fig. 3.1 Time series data collected at regular and irregular intervals of time. https://www.influxdata.com/what-is-time-series-data/

3.2 Basic Concepts in Time Series Analysis

1. **Time series data**: Time series data is a set of observations or measurements that are taken at regular intervals over a period of time. For example, daily stock prices over a year, monthly rainfall data for a region, hourly electricity consumption data for a building, etc.
2. **Trend**: The trend refers to the long-term pattern of a time series, which can be either increasing, decreasing, or stable over time.
3. **Seasonality**: Seasonality refers to the periodic fluctuations that occur in a time series within a year or a shorter period. For example, sales of air conditioners would be high in the summer months and low in the winter months.
4. **Stationarity**: A time series is said to be stationary if the statistical properties of the series, such as mean and variance, remain constant over time. This is an important concept in time series analysis as many statistical models assume stationarity.
5. **Autocorrelation**: Autocorrelation measures the correlation between observations at different points in time. If there is a high correlation between observations at different points in time, it suggests that the time series is not random and has some structure.
6. **Moving average**: A moving average is a statistical technique that is used to smooth out short-term fluctuations in a time series. It involves taking the average of a certain number of observations over a rolling time period.
7. **Exponential smoothing**: Exponential smoothing is another smoothing technique that is used to remove noise from a time series. It assigns weights to past observations and gives more weight to recent observations.
8. **Forecasting**: Forecasting is the process of using historical data to make predictions about future values of a time series. There are many statistical models used in time series forecasting, including ARIMA, exponential smoothing, and neural networks.

3.3 Uses of Time Series

Time series analysis has a wide range of applications across various fields due to its ability to analyze and extract valuable insights from sequential data. Its versatility and ability to handle sequential data make it an indispensable tool in various domains for understanding, predicting, and optimizing time-dependent phenomena.

1. Forecasting: Time series analysis is widely used for forecasting future values of a time series. This is applicable in numerous domains such as sales forecasting, demand forecasting, stock market prediction, weather forecasting, and economic forecasting. Forecasting helps organizations make informed decisions, optimize resources, and plan for the future.
2. Trend Analysis: Time series analysis can identify and analyze long-term trends in data. This is useful for understanding patterns, identifying growth or decline in variables, and making strategic decisions. Trend analysis is commonly used in finance, economics, marketing, and social sciences.
3. Seasonal Analysis: Time series analysis helps in detecting and understanding seasonal patterns within a dataset. This is particularly useful in industries like retail, tourism, agriculture, and energy, where demand and sales vary with seasons. Seasonal analysis enables businesses to optimize operations, plan inventory, and allocate resources effectively.
4. Anomaly Detection: Time series analysis can identify anomalies or outliers in data. This is valuable in various domains such as fraud detection, network monitoring, quality control, and cybersecurity. Detecting unusual patterns helps in detecting abnormalities, potential risks, and taking timely corrective actions.
5. Pattern Recognition: Time series analysis techniques, such as autocorrelation and spectral analysis, can reveal hidden patterns and dependencies within a dataset. This is beneficial in fields like signal processing, audio and speech recognition, image processing, and pattern recognition in general.
6. Decision-Making and Strategy Development: Time series analysis provides valuable insights for decision-making and strategy development. It helps businesses and organizations understand historical patterns, make informed choices, optimize processes, and develop effective strategies to achieve their goals.
7. Econometrics and Financial Analysis: Time series analysis plays a crucial role in econometrics and financial analysis. It helps economists, financial analysts, and policymakers analyze economic indicators, stock market data, interest rates, exchange rates, and other financial variables. Time series models assist in understanding the relationships between variables, predicting financial market trends, and assessing risk.
8. Process Monitoring and Control: Time series analysis is used for monitoring and controlling processes in manufacturing, engineering, and other industries. By analyzing time-dependent data, it helps in process optimization, identifying deviations, maintaining quality standards, and ensuring efficient operations.

9. Analyse: The best example of this is, the performance appraisal program that the HR department conducts in a company. When we analyse data based on the time, we can decipher the behavior of the variable. Other examples are, Chemical reactions, simulations, output from production units, etc. Statistical Quality Control (S.Q.C) is one branch that analyses the performances of production units over time.
10. Forecast: This is predominantly why we use time series analysis. In a lot of studies, time series and forecasting are paired along. With time series, we can identify approximate indicators that enable us to forecast the behavior of data with some prior information.
11. Compare: By comparing two or more datasets analyzed based on time, we understand the performance, factors contributing to the difference, and many other aspects. Vital Statistics is one branch of statistics, that deals with the population and life events of living beings. In this study, comparisons are made between cohorts (Fig. 3.2).

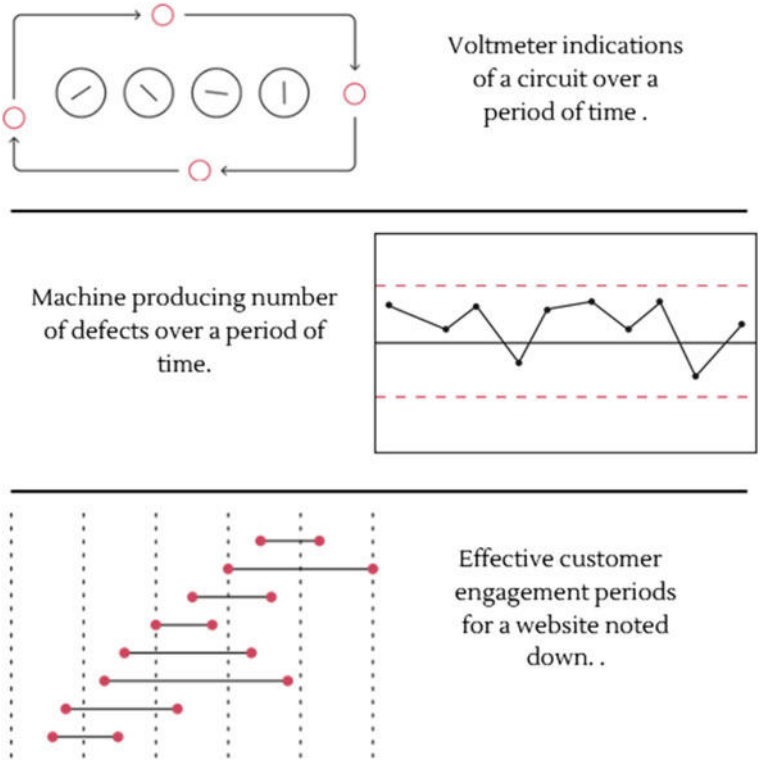


Fig. 3.2 Different ways of representing time series data

3.4 Mathematical Models of Time Series

Model building and forecasting are usually carried out based on the assumption that the data is by default distributed normally. These are the two popularly used methods, where the components of the time series are either simply added together or multiplied together based on the requirement of analysis.

(a) Additive model

$$Y = T + S + C + I$$

(b) Multiplicative model

$$Y = T * S * C * I$$

Y refers to the original time series data, T =Trend, S =Seasonal index, C =Cyclic variation and I =Irregular variations.

Let us take a visual look at the components of a time series plot for better understanding. In this example, we will also understand the additive and multiplicative preprocessing techniques in a better way. (Content taken from the site: https://www.encora.com/insights/a-visual-guide-to-time-series-decomposition-analysis) (Fig. 3.3).

Consider the dataset of 144 monthly observations from the "international airline passengers" time series dataset for the year 1949–1960.

One predominant use case of time series data is for forecasting, before we predict the data, we shall try to decompose the data for deeper understanding, the additive

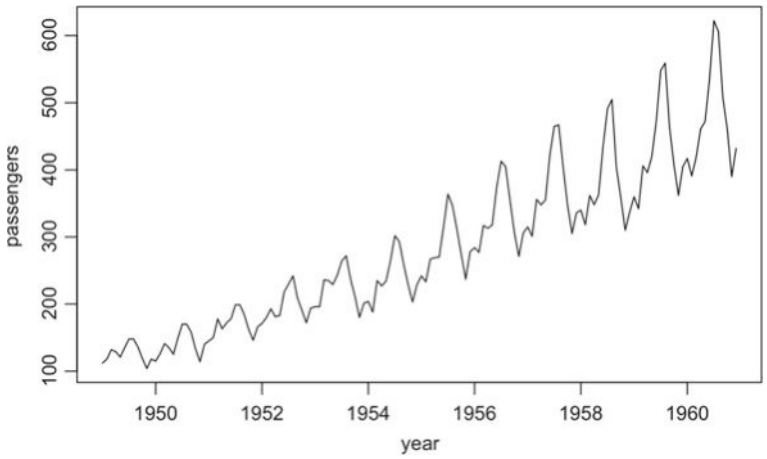


Fig. 3.3 Time series data set for the year 1949–1960

and multiplicative model refers to the different ways in which these elements/patterns of time series are combined. (These concepts are dealt with later in the chapter).

Let us consider the Additive model first:

Using the moving average method to smoothen the data, we extract the trend component of the time series using the moving averages method (which is explained later in this chapter) in order to understand the direction of the movement of data.

In Fig. 3.4 we can see the trendline across the data. Now when we eliminate the trend from the data we will be left with seasonal, cyclic and random variations, in such case, the graph looks like the one shown in Fig. 3.5. In the graph depicting the detrended values, the seasonal variations are very prominent and thus can be extracted using averaging monthly values. In several scenarios, the trend and seasonality are the two predominant components of time series Seasonality plot: In Fig. 3.6, we have seasonality plot with seasonal component.

Hence what is now left is the cyclic variation and the residual/error component in the time series. Cyclic variations can be ignored for this specific time series data as there is no strong evidence of a long-term pattern of the data. Therefore, post extraction of seasonality component, the time series with just the residual terms will be as shown in Figs. 3.7 and 3.8.

Finally, let's see the additive model overview by putting together all the graphs on a single timeline. There is loss of data at the ends of the time series due to the extraction process.

Disadvantages of additive model:

The basic idea behind the extraction of various components of time series is to understand the nature of the factors affecting the data. In additive model when these factors are clubbed together after extraction slightly varies from the original data. Let us check what we just said. In Fig. 3.9, the difference between the original data

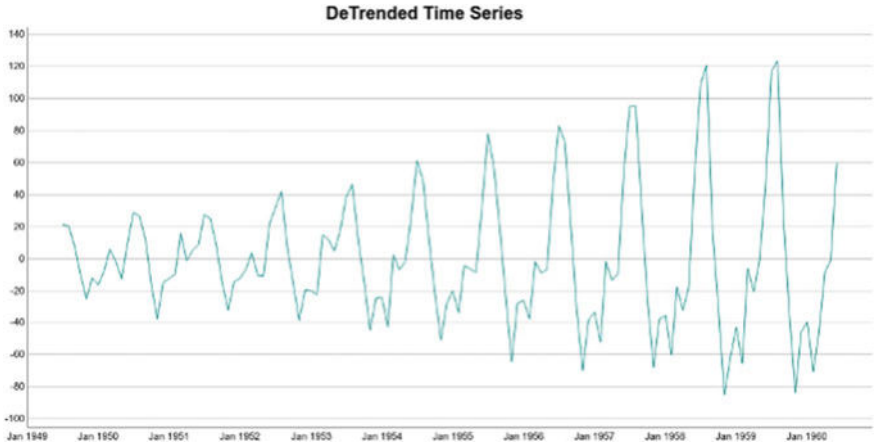


Fig. 3.4 Trendline for the time series data

Fig. 3.5 Detrended time series data

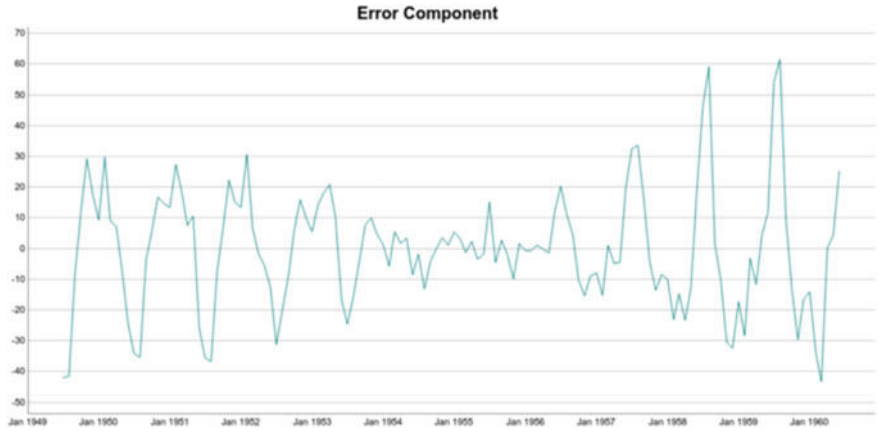


Fig. 3.6 Seasonality plot with seasonal component

and the data with trend and seasonality extracted values which are added together (additive model).

In order to significantly bridge the gap between the original data and the extracted components, a multiplicative model of time series is better than that of additive model. Take a note at figure for a better understanding.

The disadvantage that is associated with the additive model of time series data is there is a loss of data at the beginning and at the end of the series when we extract the trend line. Also during seasonal decomposition, the method assumes that the seasonal pattern repeats itself for all the years of study, which may not hold true for longer series of data as the patterns tend to change. There are also more robust

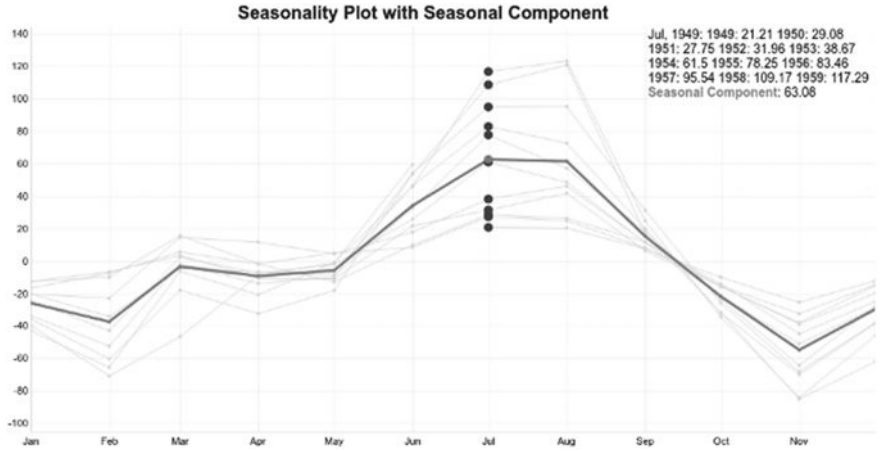


Fig. 3.7 Error component of time series

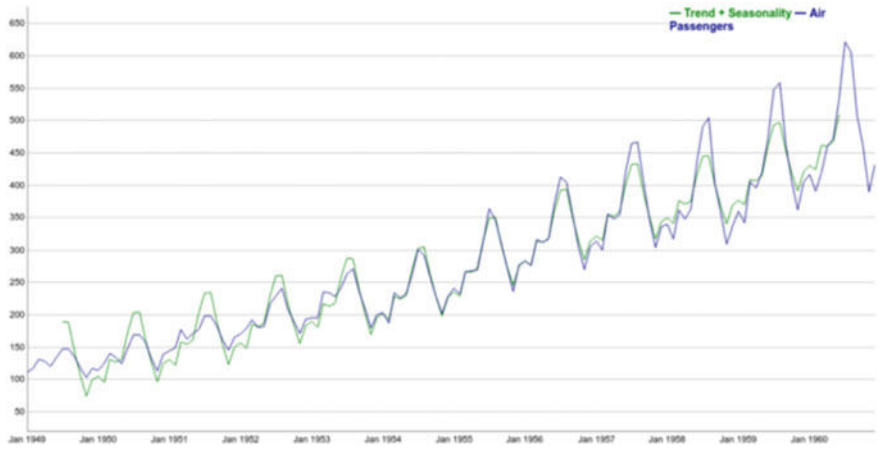


Fig. 3.8 Graph depicting all the components of time series

methods such as seasonal and rend decomposition using loess, STL decomposition and exponential smoothening and forecasting methods to overcome demerits.

3.5 Descriptive Statistics Used in Regression Analysis

Descriptive statistics are important in time series analysis to summarize and understand the characteristics of the data. Here are some commonly used descriptive statistics for time series (Fig. 3.10):

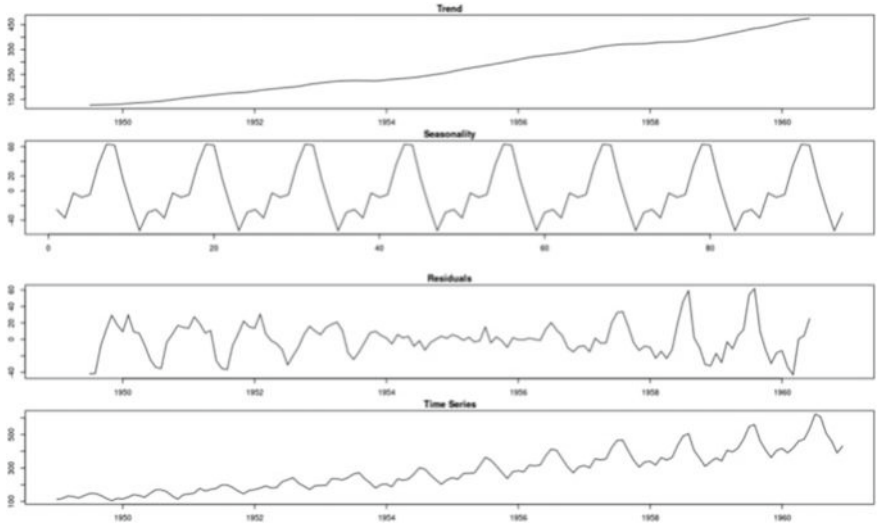


Fig. 3.9 Graph depicting trend + seasonality of time series. Additive model

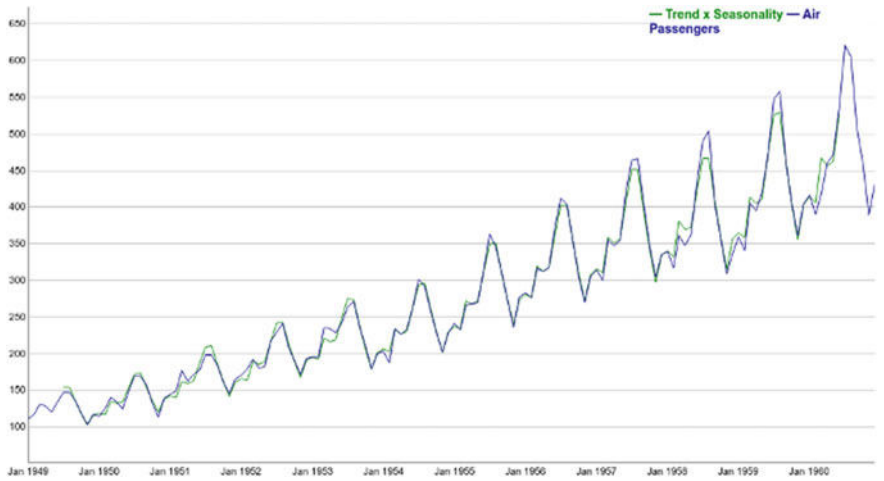


Fig. 3.10 Graph depicting trend* seasonality of time series. Multiplicative model

1. Mean: The mean or average of a time series represents the central tendency of the data. It provides information about the typical value of the series over time.
2. Variance and Standard Deviation: Variance measures the dispersion or spread of the time series data points around the mean. Standard deviation is the square root of the variance and provides a measure of the average deviation from the mean.

3. Skewness: Skewness measures the asymmetry of the distribution of the time series data. A positive skewness indicates a longer tail on the right side of the distribution, while a negative skewness indicates a longer tail on the left side.
4. Kurtosis: Kurtosis measures the peakedness or flatness of the distribution of the time series data. High kurtosis indicates a more peaked distribution with heavier tails, while low kurtosis indicates a flatter distribution.
5. Min and Max: The minimum and maximum values in a time series represent the lowest and highest observations in the dataset, providing insights into the range and extreme values.
6. Quantiles: Quantiles, such as the median (50th percentile) or quartiles (25th and 75th percentiles), divide the time series data into equal portions, providing information about the spread and central tendency of the distribution.
7. Autocorrelation: Autocorrelation measures the correlation between a time series and its lagged values. It provides insights into the temporal dependence and persistence of the series.
8. Cross-Correlation: Cross-correlation measures the correlation between two different time series at various lags. It helps understand the relationship and lagged dependencies between two variables.
9. Lagged Autocorrelation: Lagged autocorrelation coefficients, such as the partial autocorrelation function (PACF) or autocorrelation function (ACF), are used to identify the significant lagged relationships in a time series. They provide information about the lagged dependencies and potential autoregressive (AR) or moving average (MA) components in the data.

These descriptive statistics help summarize the properties, distribution, central tendency, variability, and relationships within a time series. They provide a basis for understanding the behavior and characteristics of the data, and they can be used to guide subsequent modeling, forecasting, or hypothesis testing in time series analysis.

3.6 Stationary and Non-stationary Data

Stationarity and non-stationarity are important concepts in time series analysis that describe the behavior and properties of a time series.

3.6.1 Stationarity

A time series is considered stationary when its statistical properties remain constant over time. In a stationary time series, the following conditions hold:

1. Constant Mean: The mean of the time series remains constant over time, regardless of the specific time point. It indicates that the series does not exhibit a long-term trend.

2. Constant Variance: The variance of the time series remains constant over time. It implies that the level of volatility or dispersion of the data points does not change systematically.
3. Constant Autocovariance: The autocovariance, which measures the linear relationship between the values of the time series at different time points, remains constant for all pairs of observations with the same time lag. It indicates that the correlation structure of the time series does not change over time.

Stationarity is desirable in time series analysis because it simplifies modeling and forecasting. Many statistical techniques and models assume stationarity or work better with stationary data. Stationary time series are often easier to interpret and analyze, as they exhibit consistent patterns and behaviors.

Like how a story is generally expected to start with a phrase, "Once upon a time", during time series analysis, we generally assume that the data is stationary. Example of stationary data: Heart rate captured for a normal heart from ECG.

3.6.2 Non-stationarity

A non-stationary time series, on the other hand, does not exhibit constant statistical properties over time. Non-stationarity can arise due to various reasons, including trends, seasonality, changing variances, or other time-dependent patterns. Non-stationarity makes it challenging to model and predict the behavior of the time series accurately.

Common types of non-stationarity include:

1. Trend: A time series with a systematic increase or decrease in its mean over time. Trend can be linear, quadratic, exponential, or any other form.
2. Seasonality: A repeating pattern with a fixed period, such as daily, weekly, or yearly cycles. Seasonality can cause variations in the mean and variance of the time series.
3. Changing Variance: Heteroscedasticity refers to the presence of changing variance in a time series. The level of volatility or dispersion may increase or decrease over time.
4. Unit Root: A time series with a unit root is non-stationary. It implies that the series has a stochastic trend and does not revert to a stable mean over time.

Handling Non-stationarity

Non-stationarity in a time series needs to be addressed before applying certain statistical techniques. Common approaches for handling non-stationarity include:

1. Detrending: Removing the trend component from the time series to make it stationary. This can be achieved through techniques like differencing or using regression models to model and remove the trend.

2. Seasonal Adjustment: Removing the seasonal component from the time series using methods such as seasonal differencing or seasonal decomposition of time series.
3. Transformation**:** Applying mathematical transformations, such as logarithmic transformation or power transformations, to stabilize the variance and make the series stationary.
4. Differencing: Taking differences between consecutive observations (first-order differencing, second-order differencing, etc.) to eliminate trends or seasonality.

By addressing non-stationarity, a time series can be transformed into a stationary form, making it more amenable to traditional time series models and statistical techniques.

3.7 Linear and Non-linear Time Series

Consider a data point X_t in a time series. If this data point is a linear combination of past, future and error terms are called linear time series.

$$X_t = X_{t-1} + X_{t+1} + error term.$$

Otherwise, it is called non-linear time series. Non-linear time series are of different variety of curves that a non-linear regression model can fit. These are asymmetric and also dynamic at times.

Time series data can exhibit linear or non-linear patterns and relationships. This distinction refers to the nature of the underlying behavior or dependencies within the time series. Let's explore the concepts of linear and non-linear time series:

3.7.1 Linear Time Series

A linear time series is one in which the relationship between the observations and their past values can be described by linear equations. In other words, the behavior of the time series follows a linear trend, and the impact of past values on future values is linearly related. Linear time series are relatively straightforward to model and analyze using linear regression-based methods.

Common examples of linear time series models include:

1. Autoregressive (AR) Model: An AR model expresses the current value of a time series as a linear combination of its past values and a white noise error term. The relationship between the current value and lagged values is captured by coefficients.

2. Moving Average (MA) Model: An MA model represents the current value as a linear combination of past error terms and a white noise error term. The model considers the impact of previous error terms on the current value.
3. Autoregressive Moving Average (ARMA) Model: The ARMA model combines both AR and MA components. It captures the linear relationship between the current value, past values, and previous error terms.

Linear time series models assume that the relationships between variables are constant and linear over time. They are widely used for forecasting and modeling various economic, financial, and business time series.

3.7.2 Non-linear Time Series

In contrast, non-linear time series exhibit relationships or patterns that are not adequately described by linear equations. Non-linear time series often involve complex dynamics, interactions, or dependencies between variables that cannot be captured by linear models.

Non-linear time series models are more flexible and can capture various types of non-linear behavior, such as exponential growth, oscillations, or non-linear trends. Some examples of non-linear time series models include:

1. Non-linear Autoregressive (NAR) Model: NAR models capture non-linear relationships between the current value and its lagged values, often using functions like polynomials or trigonometric functions.
2. Non-linear Moving Average (NMA) Model: NMA models incorporate non-linear dependencies between past error terms and the current value.
3. Non-linear Autoregressive Moving Average (NARMA) Model: NARMA models combine non-linear autoregressive and moving average components to capture non-linear dependencies in the time series.

Non-linear time series models require more complex estimation techniques and can be computationally intensive. They are often used when the underlying data or phenomena exhibit non-linear behavior or when linear models fail to capture important dynamics.

It's important to note that the distinction between linear and non-linear time series is not always clear-cut. In practice, the linearity assumption can be tested using statistical tests, and more advanced models like non-linear autoregressive integrated moving average (NARIMA) or machine learning algorithms can be employed to capture non-linear behavior in time series data (Fig. 3.11).

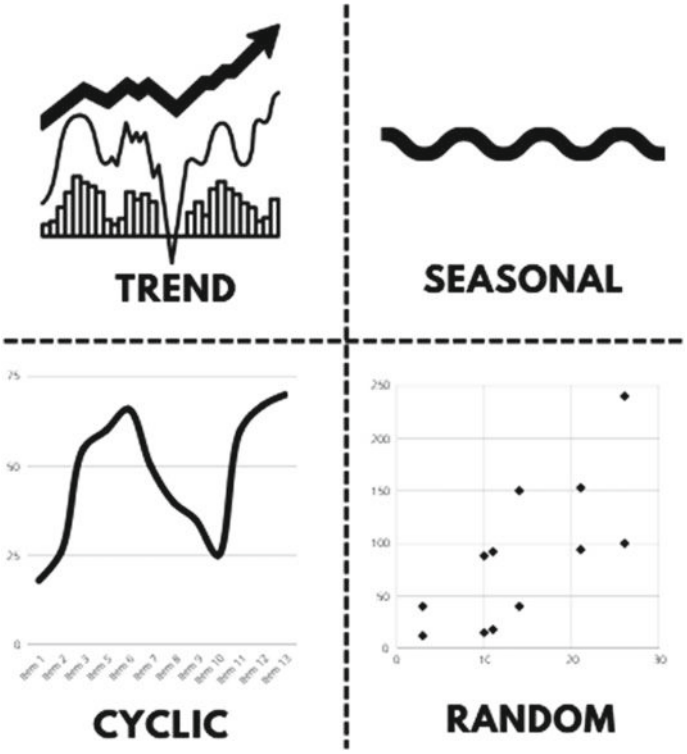


Fig. 3.11 Pictorial representation of all 4 components of time series

3.8 Components of Time Series

(a) Trend

- Refers to the general tendency of data to either increase or decrease or stagnate over a defined period of time.
- We must note that the general increase or decrease need not be consistent.
- It is often called secular trend.

(b) Seasonal variations

- Seasonal variations in time series are a component of time series in which there is a rhythmic short-term pattern created periodically (every season).
- In a graph/chart, seasonal variations are recognized as repeated rise and fall patterns across a time frame.

(c) Cyclic Variations

- Cyclic variations are also oscillatory movements where the period of oscillation is greater than one year.

- These oscillations are not uniform and not necessary to be of equal intervals of time too.

(d) Random/Irregular variations

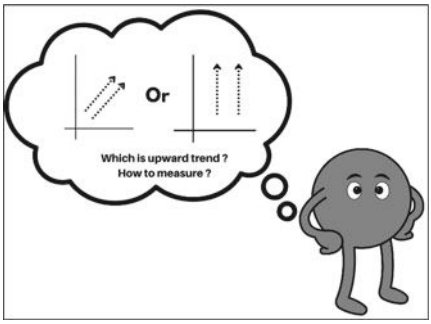
- As the name suggests, this element of time series are referred to as irregular fluctuations without a definite pattern.
- These are unpredictable, purely random and accidental changes.
- These variations are generally short-term.

3.8.1 Trend/Secular Trend

Methods of measurement of Trend

General tendency of data to either increase or decrease is referred to as TREND. There are different ways of determining the trend which are

1. Freehand or Graphical Method
2. Method of Semi-Averages
3. Method of moving averages
4. Method of Least Squares.



Freehand or Graphical Method

The flexible and easiest method of estimating the trend. The process is to first draw a histogram and then trace the flow of data using a line trying to accurately reflect the long-term tendency of data.

Advantages:

- This is a simple method for getting an overview of data.
- It does not include any mathematical procedures.

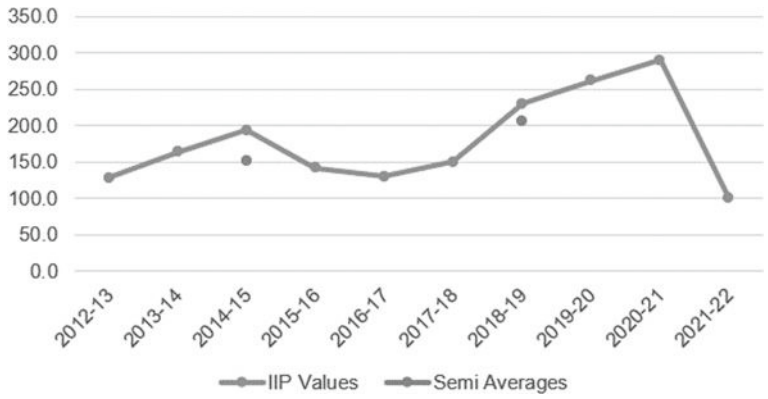


Fig. 3.12 Annual indices of IIP for primary goods

Disadvantages:

- Sometimes involves personal bias and judgment of the investigator handling the data.
- Not suitable for in-depth analysis and forecasting.

Method of Semi-Averages

A slightly more meaningful method compared to freehand method. Here we calculate the average of certain period of time (Semi-Average) and then join the dots to make an accurate trend line.

Example:

Consider the annual indices of industrial production as per use-based classification. The graph below represents the trend line that can be drawn using the two semi-average values of 5 period duration (Fig. 3.12).

Calculation:

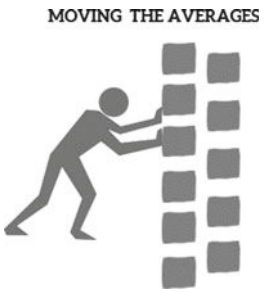
Average of I.I.P Values from 2012–2017 $= \frac{128+164+194+142+130}{5} =$ **151.6**

Average of IIP Values from 2017–2022 $= \frac{150+230+262+290+100}{5} =$ **206.4**

Method of Moving Averages

- The main idea behind moving averages is to smoothen the noise in the data for a specific period of time.
- To reduce random variations.

- The higher the period the more the lag. The analyst must decide on the best time frame that serves the purpose.
- For financial investors, to understand the support and resistance of stock, 50-day to 200-day moving averages are calculated. Whereas daily stock traders use shorter look-back periods such 2-day to 5-day moving averages for their short-term trading purposes.



Note: M.A.C.D
https://ecgtrade.com/what-is-macd-indicator-strategy.html
MACD is an acronym for Moving Average Convergence Divergence.

- This tool is used to identify moving averages that are indicating a new trend, whether it's bullish or bearish. After all, our top priority in trading is being able to find a trend, because that is where the most money is made.
- The first is the number of periods that is used to calculate the faster moving average.
- The second is the number of periods that is used in the slower moving average.
- And the third is the number of bars that is used to calculate the moving average of the difference between the faster and slower moving averages. For example, if you were to see "12, 26, 9" as the MACD parameters (which is usually the default setting for most charting packages), this is how you would interpret it
- The 12 represents the previous 12 bars of the faster moving average.
- The 26 represents the previous 26 bars of the slower moving average.
- The 9 represents the previous 9 bars of the difference between the two moving averages. This is plotted by vertical lines called a histogram (the green lines in the chart above) (Fig. 3.13).

Calculation of Moving Averages

For Odd number of years: Method of calculation for a calculation of moving average with odd number of years. Like 3, 5 and 7.

Here is an example of how calculations are made for a period of 3 years (Table 3.1, 3.2 and 3.3).

Example: 3 yearly moving average

Fig. 3.13 Moving average convergence and divergence graph

Table 3.1 Example for calculating 3 year moving average calculation

Year	Values	3 year moving total	3 year moving average
2001	a	–	–
2002	b	a +b +c =g	g/3
2003	c	b +c +d =h	h/3
2004	d	c +d +e =i	i/3
2005	e	d +e +f =j	j/3
2006	f	–	–

Table 3.2 Number for registrations received by a government school

Year	Number of registrations received	3-year moving total	3-year moving average
1985	200	–	–
1986	400	1400	466.67
1987	800	2200	733.33
1988	1000	2600	866.67
1989	800	2750	916.67
1990	950	2950	983.33
1991	1200	3150	1050.00
1992	1000	3600	1200.00
1993	1400	3900	1300.00
1994	1500	–	–

Table 3.3 Example for 5 yearly moving average

Year	Value	5 Yearly moving total	5 Yearly moving average
1988	687	–	–
1989	656	–	–
1990	639	3450	690
1991	712	3418	683.6
1992	756	3440	688
1993	655	3565	713
1994	678	3476	695.2
1995	764	3440	688
1996	623	–	–
1997	720	–	–

Consider the time series data of number of registrations received by a government school. Calculate the 3-year moving average for this data.

Example: 5 yearly moving average.

For even number of years:

With even number of years is similar to that of odd number of years, but another step for centering the average is added to position the average. Here is an example of how calculations are made for a period of 4 years, where the moving average are w, x, y, and z. There are 4 data points lost in this process. 2 at the beginning and 2 at the end.

Example: 4 yearly moving average (Table 3.4).

Method of Least Squares

The best and most precise method of calculating trend line.

Conditions to be followed:

Rule 1: The sum of deviations from the arithmetic average is zero. $\Sigma(Y-Y_c) = 0$.

where Y is the mean and Y_c is the deviation from Y. The overall difference when summed up is 0.

Rule 2: $\Sigma(Y - Y_c)^2$ is minimum.

The sum of the square of deviations from the actual and the computed value of Y is the least.

Fitting of a trend line using the linear method: $\sum XY = a \sum X + b \sum X^2 + c \sum X^3$.

Original Equation: $Y = a + b*(X)165 = a.0 + 700b + c.0$

Adding summation to all terms we get, $700b = 165$.

Equation 1: $\Sigma Y = n*a + b*(\Sigma X)$ $b = \frac{165}{700} = \mathbf{0.24}$.

Adding summation of x to all the terms to equation 1, we get,

Table 3.4 Example for 4 yearly moving average

Year	Values	4 Yearly moving total	4 Yearly moving average	4 Yearly centered moving average
1960	530	–	–	–
1961	390	–	–	–
		1851	462.75	
1962	460			460.00
		1829	457.25	
1963	471			465.75
		1897	474.25	
1964	508			493.38
		2050	512.5	
1965	458			514.88
		2069	517.25	
1966	613			520.00
		2091	522.75	
1967	490			550.50
		2313	578.25	
1968	530	–	–	–
1969	680	–	–	–

Equation 2: $\Sigma XY = a * (\Sigma X) + b * \Sigma X^2$.

Example: Given below are the figures for rice production (in lakh kg.) by a farmer. He claims that his production is profitable despite the fluctuations. His son who is an agronomist disagrees with his father and says that soil degradation with excessive of fertilizers is the reason for poor production. Plot the trend line and comment on the data available. Also, estimate the production for the year 1982. The farmer is positive about the increase in production, and his son disagrees with him (Tables 3.5 and 3.6).

Solution

Now,

$$Y_c = a + bX$$

Table 3.5 Rice produced by a farmer from 1991 to 1999

Year	1991	1992	1993	1994	1995	1996	1997	1998	1999
Value	300	450	700	720	650	675	700	950	1000

Table 3.6 Table to calculate trend line for the production of rice by farmer

Year	Production (Quintals)	Deviation from 1995			Trend value	Short-term fluctuations
	Y	X	X^2	XY	Y_c	$Y-Y_c$
1991	300	−4	16	−1200	399.1	−99.1
1992	450	−3	9	−1350	470.02	−20.02
1993	700	−2	4	−1400	540.94	159.06
1994	720	−1	1	−720	611.86	108.14
1995	650	0	0	0	682.78	−32.78
1996	675	1	1	675	753.7	−78.7
1997	700	2	4	1400	824.62	−124.62
1998	950	3	9	2850	895.54	54.46
1999	1000	4	16	4000	966.46	33.54
Total of 9 Years	6145	0	60	4255		0

$$a = \frac{\sum y}{N} = \frac{6145}{9} = 682.78$$

$$b = \frac{\sum xy}{\sum x^2} = \frac{4255}{60} = \mathbf{70.92}$$

Equation of the trend line $= \mathbf{Y_c = 682.78 + 70.92X}$.

We can notice that the total of short-term fluctuations column that has the values of $Y-Y_c$ is 0. Which again means that the total sum of deviations around the mean is zero.

Note: Fitting of a quadratic equation needs changes in both the regression equation as well as the normal equation.

Example: The table below shows the production of A2 pasteurized cow milk from the main branch of a milk factory. Fit a second-degree parabola for the data given below (Table 3.7).

$$\boldsymbol{Y_c = a + bX + cX^2}$$

Table 3.7 Production of A2 pasteurized cow milk from the main branch of a milk factory

Year	1975	1980	1985	1990	1995	2000	2005
Production	12	14	10	11	18	16	19

Solution

The equation for a parabola trend line =

The values of a,b, and c can be obtained by solving the following equations,

$$\sum Y = Na + b \sum X + c \sum X^2$$

$$100 = 7a + b.0 + 700c$$

$7a + 700c = 100$ —Equation 1

$$\sum X^2 Y = a \sum X^2 + b \sum X^3 + c \sum X^4$$

$$10{,}675 = 700a + b.0 + 122{,}500c$$

$$700a + 122{,}500c = 10{,}675 \text{ –Equation 2}$$

Solving for variables a and c we get (Table 3.8),

$$a = \frac{\sum Y - c \sum X^2}{N} = \frac{100 - (0.012)(700)}{7} = 13$$

$$b = \frac{\sum XY}{\sum X^2} = \frac{165}{700} = 0.24$$

$$c = \frac{N \sum X^2 Y - \sum X^2 \sum Y}{N \sum X^4 - (\sum X^2)^2} = \frac{(7)(106{,}75) - (700)(100)}{(7)(122{,}500) - (700)^2} = 0.012$$

Therefore, the equation of the trend line is,

Table 3.8 Table representing the calculations to estimate the production of rice for the year 2010

Year	Production units in thousands (Y)		Calculations					Trend Values Y_c
		Deviation from 1990 (X)	X^2	X^3	X^4	XY	X^2Y	
1975	12	−15	225	−3375	50,625	−180	2700	12.10
1980	14	−10	100	−1000	10,000	−140	1400	11.80
1985	10	−5	25	−125	625	−50	250	12.10
1990	11	0	0	0	0	0	0	13.00
1995	18	5	25	125	625	90	450	14.50
2000	16	10	100	1000	10,000	160	1600	16.60
2005	19	15	225	3375	50,625	285	4275	19.30
N =7	100	0	700	0	122,500	165	10,675	

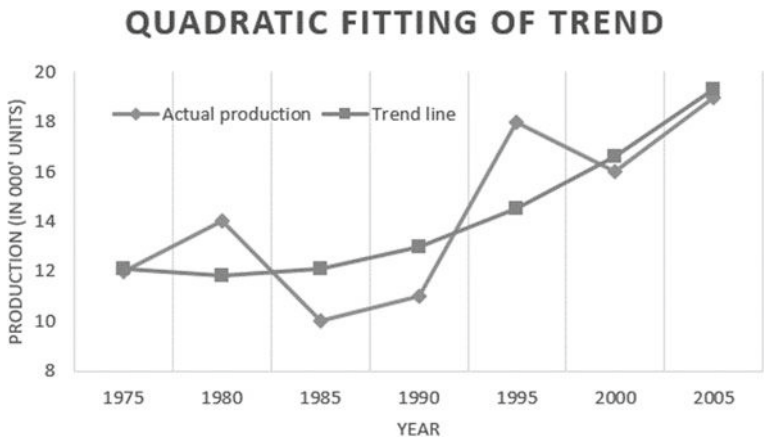


Fig. 3.14 Quadratic fitting of trend for production of rice

$$\mathbf{Y_c} = 13 + 0.24\mathbf{X} + 0.012\mathbf{X}^{\ 2}$$

The estimate for year 2010 is given by (Fig. 3.14)

$$Y_{2010} = 13 + (0.24)(20) + (0.012)(400) = 22.6$$

3.9 Seasonal Variations

Seasonal fluctuations in time series refer to the regular/periodic fluctuations in the time series that are less than a period of one year. The elimination of seasonal variation from the time series is called deseasonalization. Deriving the seasonal fluctuations from a time series.

1. Additive model: S = Y – (T + C + I)
2. Multiplicative model: $S = \frac{Y}{T*C*I} * 100$

We can usually identify an additive or multiplicative time series from its variation. If the magnitude of the seasonal component changes with time, then the series is multiplicative. Otherwise, the series is additive. Notice that the magnitude of the seasonal component—the difference between the maximum point of the series and the red line is relatively constant from 2011 onward in the additive model (Fig. 3.15).

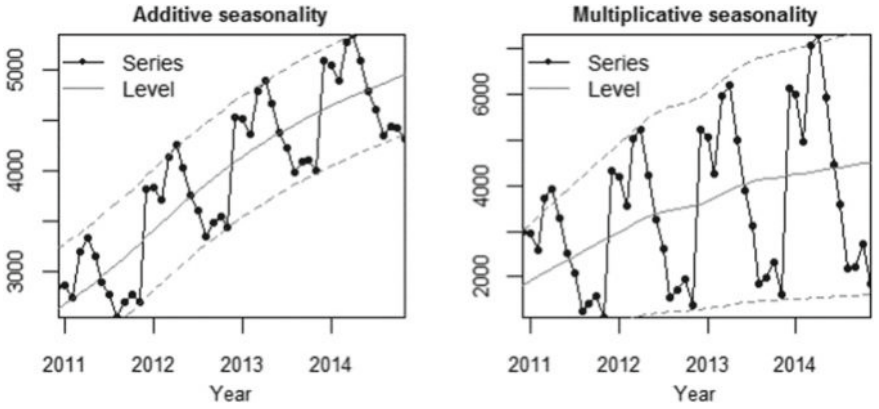


Fig. 3.15 Additive and multiplicative model of seasonal variation. Image from Nikolaos Kourentzes

3.9.1 Methods of De-Seasonalizing Data

1. Method of simple averages.
2. Ratio to trend method.
3. Ratio to moving average method.

Method of Simple Averages

Description:

For this method, we require the data that is divided into a defined period of time, either quarterly/monthly/weekly/hourly, etc. This is one of the crude and simplest methods extracting the seasonal effect of time series.

Merits and demerits:

This method assumes that the time series data is free from trend and cyclic variations. It averages out seasonal fluctuations and reduces the effect of irregularities. This makes the process simple but reduces its practicality with respect to real-life applications (Tables 3.9 and 3.10).

Example:

$$\text{Grand average} = \frac{\text{Sum of 4 Quarterly Averages}}{4} = \frac{72.6 + 71.4 + 72 + 72.4}{4}$$
$$= \frac{288.4}{4} = 72.1$$

Calculation of seasonal indices

Table 3.9 Quarterly data to understand de-seasonalizing data by the method of simple average

Years	I Quarter	II Quarter	III Quarter	IV Quarter
2008	70	65	60	75
2009	75	70	75	70
201	72	75	70	72
2011	76	72	75	75
2012	70	75	80	70

Table 3.10 Calculation of seasonal indices by method of simple average

Years	I Quarter	II Quarter	III Quarter	IV Quarter
2008	70	65	60	75
2009	75	70	75	70
2010	72	75	70	72
2011	76	72	75	75
2012	70	75	80	70
Quarterly total	363	357	360	362
Quarterly average	72.6	71.4	72	72.4
Seasonal indices	100.69	99.03	99.86	100.42

$$\text{S.I for I Quarter} = \frac{\text{Average of I Quarter}}{\text{Grand Average}} * 100 = \frac{72.6}{72.1} * 100 = 100.69$$

$$\text{S.I for II Quarter} = \frac{\text{Average of II Quarter}}{\text{Grand Average}} * 100 = \frac{71.4}{72.1} * 100 = 99.03$$

$$\text{S.I for III Quarter} = \frac{\text{Average of III Quarter}}{\text{Grand Average}} * 100 = \frac{72}{72.1} * 100 = 99.86$$

$$\text{S.I for IV Quarter} = \frac{\text{Average of IV Quarter}}{\text{Grand Average}} * 100 = \frac{72.4}{72.1} * 100 = 100.42$$

Ratio to Trend Method

This method provides seasonal indices free from trend and is an improved version of the simple average method as it assumes that seasonal variation for a given period is a constant fraction of the trend.

Merits and demerits:

In this method, the process assumes that the seasonal variations in time series are a factor of the trend values. Therefore, in the process of extraction, the original values

are expressed in terms of the percentage of the trend values. Though this method is an improvement over the first method of deseasonalization. The practical usage of this method is still limited as they ignore the cyclic effect of time series data. This method also holds an advantage of no-loss-data over the "ratio to moving averages method."

Problem: Quarterly purchases of anesthesia doses for a hospital is listed below, derive the seasonal pattern from the data and check if we can deduce any information from it (Table 3.11).

Solution

Computation of trend (Tables 3.12 and 3.13).

Trend values for Quarter 1:

$$a_1 = \frac{\sum Y_1}{N} = \frac{224}{5} = 44.8 \quad b_1 = \frac{\sum XY_1}{\sum X^2} = \frac{68}{10} = 6.8$$

$$\mathbf{Y_{c_1} = 44.8 + 6.8X}$$

$$\mathbf{Y_{c_2} = a_2 + b_2 X}$$

Trend values for Quarter 2:

Table 3.11 Quarterly data to understand de-seasonalizing data by the method of ratio to trend

Year	1st Quarter	2st Quarter	3st Quarter	4st Quarter
2011	40	46	38	52
2012	32	48	42	46
2013	36	36	54	62
2014	52	68	76	72
2015	64	50	74	88

Table 3.12 Calculation of seasonal indices by ratio to trend method

Year	Quarter 1	Quarter 2	Quarter 3	Quarter 4	X Deviation from 2013	X^2	XY_1	XY_2	XY_3	XY_4
	Y_1	Y_2	Y_3	Y_4						
2011	40	46	38	52	−2	4	−80	−92	−76	−104
2012	32	48	42	46	−1	1	−32	−48	−42	−46
2013	36	36	54	62	0	0	0	0	0	0
2014	52	68	76	72	1	1	52	68	76	72
2015	64	50	74	88	2	4	128	100	148	176
N = 5	224	248	284	320	0	10	68	28	106	98

Table 3.13 Table representing seasonal variations and trend values

Year	Trend values			
	Y_{C1}	Y_{C2}	Y_{C3}	Y_{C4}
2011	31.2	44	35.6	44.4
2012	38	46.8	46.2	54.2
2013	44.8	49.6	56.8	64
2014	51.6	52.4	67.4	73.8
2015	58.4	55.2	78	83.6
Year	Seasonal variations			
	Quarter 1	Quarter 2	Quarter 3	Quarter 4
2011	78	95.65	93.68	85.38
2012	118.75	97.5	110	117.83
2013	124.44	137.78	105.19	103.23
2014	99.23	77.06	88.68	102.5
2015	91.25	110.4	105.41	95

$$a_2 = \frac{\sum Y_2}{N} = \frac{248}{5} = 49.6 \quad b_2 = \frac{\sum XY_2}{\sum X^2} = \frac{28}{10} = 2.8$$

$$\mathbf{Y_{c_2} = 49.6 + 2.8X}$$

Trend values for Quarter 3: $Y_{c_3} = a_3 + b_3 X$

$$a_3 = \frac{\sum Y_3}{N} = \frac{284}{5} = 56.8 \quad b_3 = \frac{\sum XY_3}{\sum X^2} = \frac{106}{10} = 10.6$$

$$\mathbf{Y_{c_3} = 56.8 + 10.6X}$$

Trend values for Quarter 4: $Y_{c_4} = a_4 + b_4 X \; Y_{c_4} = a_4 + b_4 X$

$$a_4 = \frac{\sum Y_4}{N} = \frac{320}{5} = 64 \quad b_4 = \frac{\sum XY_4}{\sum X^2} = \frac{98}{10} = 9.8$$

$$Y_{c_4} = 64 + 9.8X$$

Computing the seasonal indices
Variation for quarter 1 values $= \frac{Y_{c_1}}{Y_1} * 100$.
For year 2011 $= \frac{31.2}{40} = 78$ and so on
Variation for quarter 2 values $= \frac{Y_{c_2}}{Y_2} * 100$.
For year 2011,$= \frac{46}{44} = 95.65$ and so on.
Variation for quarter 3 values $= \frac{Y_{c_3}}{Y_3} * 100$.
For year 2011 $= \frac{38}{35.6} = 93.68$ and so on.

Variation for quarter 3 values $= \frac{Y_{c3}}{Y_3} * 100$.
For year 2011 $= \frac{52}{44.40} = 85.38$ and so on.

3.9.2 Ratio to Moving Averages Method

The ratio to moving averages (RMA) method is a technical analysis tool used to identify trends and potential buy or sell signals in financial markets. It involves calculating the ratio of two moving averages of an asset's price, typically a short-term moving average and a long-term moving average. The RMA method is based on the idea that crossovers between these moving averages can indicate shifts in the market's momentum.

Suppose we have the following daily closing prices of a stock over a 10-day period:

Day 1: \$50.
Day 2: \$52.
Day 3: \$55.
Day 4: \$58.
Day 5: \$54.
Day 6: \$53.
Day 7: \$49.
Day 8: \$47.
Day 9: \$48.
Day 10: \$52.

We'll calculate two moving averages: a short-term moving average (SMA) over 5 days and a long-term moving average (LMA) over 10 days.

1. Calculate the short-term moving average (SMA):

$$\begin{aligned}
&\text{SMA} = (\text{Day 1} + \text{Day 2} + \text{Day 3} + \text{Day 4} + \text{Day 5}\quad)/5\\
&\text{SMA} = (50 + 52 + 55 + 58 + 54\quad)/5\\
&\text{SMA} = 269/5\\
&\text{SMA} = 53.8
\end{aligned}$$

2. Calculate the long-term moving average (LMA):

$$\begin{aligned}
&\text{LMA} = (\text{Day 1} + \text{Day 2} + \text{Day 3} + \text{Day 4} + \text{Day 5} + \text{Day 6} + \text{Day 7} + \text{Day 8}\\
&\quad + \text{Day 9} + \text{Day 10}\,)/10\\
&\text{LMA} = (50 + 52 + 55 + 58 + 54 + 53 + 49 + 47 + 48 + 52\qquad)/10\\
&\text{LMA} = 518/10\\
&\text{LMA} = 51.8
\end{aligned}$$

Now that we have calculated both the short-term SMA and the long-term LMA, we can find the ratio:

$$\begin{aligned}\text{Ratio} &= \text{SMA}/\text{LMA}\\ \text{Ratio} &= 53.8/51.8\\ \text{Ratio} &\approx 1.038\end{aligned}$$

The ratio we obtained is approximately 1.038. In the RMA method, traders and analysts often use specific threshold values to generate buy or sell signals. If the ratio crosses above a certain threshold (e.g., 1), it may indicate a bullish trend and a potential buy signal. Conversely, if the ratio falls below a threshold (e.g., 1), it may indicate a bearish trend and a potential sell signal.

Keep in mind that this is a simplified example, and in real-world scenarios, traders often use more sophisticated techniques, additional indicators, and historical data to make well-informed decisions. Technical analysis tools like the RMA method should be used in conjunction with other methods and risk management strategies for successful trading.

3.10 Time Series and Stochastic Processes

Time series analysis is a statistical method used to analyze and interpret data that is collected over time. In time series, observations are recorded in a sequential order at regular intervals, such as hourly, daily, monthly, or yearly. Examples of time series data include stock prices, temperature readings, and monthly sales figures.

Stochastic processes, on the other hand, are mathematical models that describe the evolution of a system over time. They are used to model situations where there is inherent randomness or uncertainty in the underlying processes. Stochastic processes can be used to generate time series data that exhibit specific statistical properties.

There are different types of stochastic processes, and one common type is the Markov process. A Markov process is a sequence of random variables where the probability distribution of each variable depends only on the previous variable in the sequence. This property is known as the Markov property or memorylessness.

Another important concept in time series analysis is stationarity. A stationary time series is one whose statistical properties, such as mean and variance, do not change over time. This assumption is often made when applying various statistical techniques to analyze time series data.

Time series analysis involves several key steps. First, exploratory data analysis is performed to understand the patterns and characteristics of the data. This includes examining plots, calculating summary statistics, and checking for trends, seasonality, and outliers.

Next, various techniques are applied to model and forecast the time series. These techniques can include autoregressive integrated moving average (ARIMA) models,

which capture both the autoregressive and moving average components of the time series. Other approaches include exponential smoothing models, state space models, and machine learning methods such as recurrent neural networks (RNNs) or long short-term memory (LSTM) networks.

Once a model is selected, it can be used for forecasting future values or analyzing the underlying patterns and relationships in the data. Model diagnostics and validation are important to ensure the model's adequacy and reliability.

Thus, time series analysis and stochastic processes are important tools for understanding and modeling data that evolves over time. They provide insights into the patterns, trends, and behavior of time-dependent data, enabling us to make predictions, detect anomalies, and uncover underlying relationships.

3.10.1 Difference Between Time Series and Stochastic Process

While time series and stochastic processes are related concepts, they have some key differences.

1. Definition and Focus: Time series specifically refers to a sequence of observations recorded at regular intervals over time. The focus is on analyzing and understanding the patterns, trends, and characteristics of the data in the time domain. Time series analysis aims to model, forecast, and make inferences about the future behavior of the data based on its historical patterns.

On the other hand, a stochastic process is a mathematical model that describes the evolution of a system over time. It is a broader concept that encompasses time series but is not limited to it. Stochastic processes provide a framework for modeling random phenomena, incorporating randomness, uncertainty, and dependencies. Stochastic processes can generate time series data, but they can also describe other types of random processes, such as spatial processes or random walks.

2. Mathematical Formalism: Time series analysis typically focuses on the statistical properties and modeling techniques specific to analyzing time-dependent data. It involves analyzing autocorrelation, trend estimation, seasonality, and other time-related patterns. Methods such as autoregressive integrated moving average (ARIMA) models, exponential smoothing, and spectral analysis are commonly used in time series analysis.

Stochastic processes, on the other hand, are more abstract mathematical models. They are defined in terms of probability theory and often involve concepts such as Markov chains, transition probabilities, and probability distributions. Stochastic processes provide a mathematical framework to model randomness, capture dependencies, and study the behavior of random variables over time or other dimensions.

3. Applications: Time series analysis finds applications in various fields such as finance, economics, engineering, environmental sciences, and social sciences.

It is used to analyze and forecast economic indicators, stock prices, weather patterns, and many other time-dependent phenomena.

Stochastic processes have broader applications beyond time series analysis. They are used in fields such as physics, biology, operations research, and machine learning. Stochastic processes can model phenomena such as particle movement, population dynamics, traffic flow, optimization problems, and more.

Thus, time series analysis is focused specifically on analyzing and modeling time-dependent data, while stochastic processes provide a mathematical framework for modeling randomness and dependencies in various domains beyond just time series. Time series can be seen as a specific type of stochastic process that deals with sequential data recorded over time.

3.10.2 Examples of Stochastic Processes

There are various examples of stochastic processes that are commonly used to model real-world phenomena. Here are a few examples:

1. Random Walk: A random walk is a simple example of a stochastic process. It represents a sequence of random steps taken in a random direction. Each step is determined by a random variable, and the process evolves based on the cumulative sum of these random variables. Random walks are used to model phenomena such as stock prices, population growth, and particle movement.
2. Brownian Motion: Brownian motion is a specific type of random walk where the steps are normally distributed with zero mean and constant variance. It is named after the botanist Robert Brown, who observed the erratic movement of pollen particles suspended in a liquid. Brownian motion is widely used to model various phenomena, including financial markets, diffusion processes, and thermal fluctuations.
3. Poisson Process: A Poisson process is a stochastic process that models events occurring randomly over time. It is characterized by the property that the number of events in any time interval follows a Poisson distribution. Poisson processes are used to model phenomena such as the arrival of customers at a service counter, the occurrence of earthquakes, or the arrival of photons in a digital communication system.
4. Markov Chain: A Markov chain is a stochastic process that transitions between a set of states according to a probability distribution. The probability of transitioning to a particular state depends only on the current state and not on the past history. Markov chains are widely used in various applications, including modeling weather patterns, stock market behavior, and language generation.
5. Autoregressive (AR) Process: An autoregressive process is a stochastic process where the current value depends linearly on its past values and a random error term. The term "autoregressive" refers to the fact that the process regresses

on itself. AR processes are used to model time series data with temporal dependencies, such as stock prices, economic indicators, and weather data.
6. Moving Average (MA) Process: A moving average process is a stochastic process where the current value depends linearly on the past random error terms and a random error term in the current period. It represents a weighted sum of past error terms. MA processes are commonly used in time series analysis, often in conjunction with autoregressive processes, to model and forecast data.
7. Gaussian Process: A Gaussian process is a stochastic process where any finite set of observations follows a multivariate normal distribution. It is defined by a mean function and a covariance function, which capture the prior knowledge about the process. Gaussian processes have applications in machine learning, spatial statistics, and optimization.

These examples represent just a small subset of the wide range of stochastic processes used in different fields. Each process has its own characteristics and mathematical properties, making them suitable for modeling specific types of phenomena and providing insights into the behavior of random variables over time.

3.11 What Are Lagged Values?

Lagged values, in the context of time series analysis, refer to the values of a variable at previous time points. When analyzing a time series, it is common to consider the relationship between an observation at a particular time point and its past observations.

The concept of lagged values arises from the idea that the current value of a time series may be influenced by its past values. By examining the relationship between the current value and lagged values, we can gain insights into the temporal dependencies and patterns present in the data.

The lag of a variable represents the number of time units (such as days, months, or years) by which the observation is shifted backward in time. For example, a lag of 1 indicates that the observation at time t is compared to the value at time t-1, while a lag of 2 compares the value at time t to the value at time t-2.

Lagged values are often used in time series analysis for various purposes, including:

1. Autocorrelation: Lagged values are used to calculate autocorrelation, which measures the correlation between a time series and its lagged values. Autocorrelation helps identify any temporal patterns or dependencies in the data, such as seasonality or trend.
2. Autoregressive Models: In autoregressive models, the current value of a time series is modeled as a linear combination of its past values. Lagged values are used as predictors in the model to capture the temporal dependencies. For example, an AR(1) model uses the value at the previous time step (lag 1) as a predictor.

3. Forecasting: Lagged values can be used as features in forecasting models. By including lagged values as predictors, we can leverage the information from previous time points to make predictions about future values.
4. Time Series Decomposition: Lagged values are used in time series decomposition techniques such as seasonal decomposition of time series (e.g., using seasonal and trend components). These techniques help separate the time series into its underlying components, such as trend, seasonality, and residuals.

By examining the relationship between a time series and its lagged values, we can uncover important temporal patterns, dependencies, and dynamics in the data, enabling us to make more accurate predictions and gain insights into the underlying behavior of the time series.

3.12 Graphical Representation of Time Series

There are several graphical representations commonly used to visualize time series data. These visualizations help to understand the patterns, trends, and other characteristics present in the data. Here are some of the most commonly used graphical representations of time series:

1. Time Plot: A time plot is a simple line graph where the time series observations are plotted on the y-axis against time on the x-axis. This plot shows the changes in the values of the time series over time, allowing for visual inspection of trends, seasonality, and other patterns.
2. Seasonal Plot: A seasonal plot is a variation of the time plot that focuses on the seasonal patterns in the data. It displays the data for multiple seasonal periods, typically in a single year, in separate panels or overlaid on top of each other. This plot helps to identify repeating patterns and seasonality effects.
3. Scatter Plot: A scatter plot is useful when examining the relationship between two variables in a time series. It displays individual data points as dots on a graph, with one variable on the x-axis and the other on the y-axis. This plot can reveal any correlation or patterns between the two variables.
4. Histogram: A histogram is a graphical representation that shows the distribution of values in a time series. It divides the range of values into bins and displays the frequency or count of observations falling into each bin. Histograms can provide insights into the shape of the distribution, such as whether it is symmetric, skewed, or has multiple peaks.
5. Box Plot: A box plot, also known as a box-and-whisker plot, provides a summary of the distribution of a time series. It displays the minimum, maximum, median, and quartiles of the data. This plot can help identify outliers, the spread of the data, and the presence of skewness.

6. Autocorrelation Plot: An autocorrelation plot, also called a correlogram, shows the correlation of a time series with its lagged values. It plots the autocorrelation coefficient on the y-axis against the lag on the x-axis. This plot is useful for identifying any significant lagged relationships or seasonality effects.
7. Spectral Plot: A spectral plot, often obtained through a technique called spectral analysis, provides insights into the frequency components present in a time series. It displays the power or amplitude of each frequency on the y-axis against frequency on the x-axis. This plot can be useful for detecting periodicities and dominant frequencies in the data.

These graphical representations can be created using various software tools like Python's Matplotlib, R's ggplot2, or other specialized time series analysis software. Each plot provides a different perspective on the time series data and can aid in identifying patterns, trends, seasonality, outliers, and other important features of the data.

3.13 General Overview of the Steps Involved in Time Series Data Processing

1. Data Collection: The first step in time series data processing is collecting the data. This can be done through various means such as sensors, databases, or web scraping, depending on the source of the data.

Data collection in time series refers to the process of gathering and recording data points over a specific period at regular intervals. It involves systematically capturing observations or measurements of a variable of interest at different points in time. Here are some common methods and considerations for data collection in time series:

1. Sampling Frequency: Determine the frequency at which data will be collected, such as hourly, daily, weekly, monthly, etc. The choice of frequency depends on the nature of the phenomenon being measured and the purpose of the analysis.
2. Data Sources: Identify the sources from which the time series data will be collected. This can include sensors, instruments, databases, surveys, web scraping, or any other means that provide access to the required information.
3. Data Quality: Ensure the quality and reliability of the collected data. Implement measures to minimize errors, missing values, outliers, and other data issues. Data cleaning and preprocessing techniques may be necessary to address any anomalies or inconsistencies in the collected data.
4. Data Storage and Organization: Establish a suitable data storage system and structure to efficiently manage and retrieve the collected time series data. Consider using databases, spreadsheets, or specialized time series data management tools.

5. Data Documentation: Maintain documentation that describes the data collection process, including details such as the data source, sampling frequency, measurement units, any transformations or adjustments applied, and any relevant metadata.
6. Data Security and Privacy: Implement appropriate measures to ensure data security and protect the privacy of individuals or organizations associated with the collected time series data. Adhere to relevant data protection regulations and best practices.
7. Data Validation: Validate the collected data against established criteria to ensure its accuracy and consistency. This may involve cross-checking with other data sources or using statistical techniques to identify potential issues.
8. Data Continuity: Aim for consistent and uninterrupted data collection over the desired time period. Ensure that the data collection process remains active and operational, taking into account potential disruptions, maintenance schedules, or changes in data sources.
9. Data Documentation Updates: Continuously update and maintain the documentation as new data is collected or changes occur in the data collection process. This helps in preserving the context and integrity of the time series data for future analysis.
10. Ethical Considerations: Consider ethical implications associated with the data collection process, such as informed consent, data anonymization, and adherence to ethical guidelines and regulations.

It's important to plan and execute the data collection process carefully to ensure the availability of high-quality time series data that accurately represents the phenomenon under study.

2. **Data Cleaning**: Once the data is collected, it is essential to clean it to handle missing values, outliers, and any other inconsistencies and get it ready for analysis. Steps in Data cleaning in Time series are as follows:

1. Handling Missing Values: Missing values can be problematic in time series data, as they disrupt the continuity and can affect subsequent analysis. You can handle missing values by either removing the corresponding time points or imputing them with appropriate values. Imputation methods include forward filling, backward filling, interpolation, or using advanced imputation techniques such as regression-based imputation.
2. Outlier Detection and Treatment: Outliers are extreme values that deviate significantly from the majority of the data points. Outliers can occur due to measurement errors or other anomalies. Identifying and treating outliers is essential for accurate analysis. Outliers can be detected using statistical methods such as the Z-score, percentile-based methods, or visual inspection. Treatment options include removing the outliers or transforming them to minimize their impact.
3. Handling Irregular Sampling: Time series data may sometimes have irregular sampling intervals or missing time points. In such cases, you might need to resample the data to a regular interval using techniques like interpolation or downsampling. This ensures uniformity and consistency in the time series data.

4. Addressing Seasonality and Trends: Time series data can exhibit seasonality (repeating patterns) and trends (long-term changes). It's important to identify and remove or model these components to better understand the underlying patterns and relationships. Techniques such as seasonal decomposition or detrending can help separate the seasonality and trend components from the data.
5. Data Transformation: In some cases, transforming the data can improve its properties or make it more amenable to analysis. Common transformations include logarithmic transformation, differencing, or scaling the data to a specific range. These transformations can help stabilize variance, reduce skewness, or remove trends in the data.
6. Handling Data Inconsistencies: Time series data might also suffer from inconsistencies or errors such as duplicate entries, incorrect timestamps, or inconsistent units. It's crucial to carefully check for such inconsistencies and rectify them to ensure the integrity of the data.
7. Data Normalization: Normalizing the data can be useful when working with multiple time series with different scales or units. Normalization techniques such as min–max scaling or z-score normalization can bring the data to a common scale and facilitate meaningful comparisons.

Remember that the specific data cleaning steps may vary depending on the characteristics of your time series data and the analysis objectives. It's essential to thoroughly understand the data, carefully review its quality, and apply appropriate cleaning techniques to ensure reliable and accurate results in subsequent time series analysis.

3. Data Transformation: Time series data often exhibits non-linear patterns and trends. To make the data more suitable for analysis, various transformations can be applied, such as taking logarithms, differencing, or applying Box-Cox transformations. These transformations help stabilize the variance and make the data more stationary.

Data transformation in time series refers to the process of altering the original time series data to make it more amenable for analysis or to satisfy certain assumptions of statistical models. It involves applying mathematical or statistical operations to the data to achieve specific objectives. Here are some common data transformation techniques used in time series analysis:

1. Logarithmic Transformation: Taking the logarithm of the data values is often used to stabilize the variance when the data exhibits exponential growth or decay. This transformation can be useful when the variability of the data increases with the magnitude of the values.
2. Box-Cox Transformation: The Box-Cox transformation is a power transformation that generalizes the logarithmic transformation. It allows for a wider range of transformations by introducing a parameter (lambda) that determines the type and degree of transformation applied. The optimal lambda value can be estimated through statistical techniques.

3. Difference Transformation: Differencing involves computing the differences between consecutive observations in the time series. It is commonly used to remove trends or seasonality from the data, making it more stationary. First-order differencing subtracts each value from its preceding value, while higher-order differencing can be performed if additional differencing is necessary.
4. Seasonal Difference Transformation: In the presence of seasonal patterns, seasonal differencing can be applied to remove the seasonality from the data. It involves computing the differences between observations from the same season but in different years. Seasonal difference can be combined with regular differencing for further stabilization of the data.
5. Scaling and Standardization: Scaling and standardization techniques are used to normalize the data by shifting and rescaling it. Common methods include min–max scaling, where the data is scaled to a specific range (e.g.: 0 to 1), and z-score standardization, which transforms the data to have a mean of 0 and a standard deviation of 1.
6. Smoothing Techniques: Smoothing methods, such as moving averages or exponential smoothing, can be applied to reduce noise and short-term fluctuations in the data. These techniques help reveal underlying patterns and trends by averaging out random variations.
7. Fourier Transformation: Fourier transformation is used to decompose the time series into its frequency components. This transformation is particularly useful for identifying periodic patterns or seasonality in the data.
8. Data Aggregation: Aggregating the data by combining multiple observations into larger time intervals (e.g., from hourly to daily) can help reduce noise and provide a clearer picture of the overall trends and patterns.
9. Data Discretization: Discretization involves converting continuous time series data into discrete intervals or categories. This can be useful for analyzing data in a categorical or interval-based framework.
10. Winsorization: Winsorization is a technique that replaces extreme values (outliers) in the data with less extreme values. This helps mitigate the influence of outliers on the analysis.

These are just a few examples of data transformation techniques used in time series analysis. The choice of transformation method depends on the characteristics of the data and the specific objectives of the analysis. It's important to consider the impact of the transformation on the interpretation and results of subsequent analyses.

4. Resampling: Time series data may be collected at irregular intervals or have a high-frequency resolution that is not required for the analysis. Resampling techniques such as upsampling (increasing frequency) or downsampling (decreasing frequency) can be applied to align the data with the desired time intervals.

Resampling in time series refers to the process of changing the frequency or granularity of the data by aggregating or disaggregating the original observations. It involves converting the time series data from one time scale to another, such as increasing or decreasing the frequency or changing the time intervals. Resampling is

useful for various purposes, such as aligning data to a common time frame, smoothing the data, or preparing it for different types of analyses. Here are two common types of resampling techniques used in time series analysis:

1. Upsampling (Increasing Frequency):
    - Upsampling involves increasing the frequency of the data by creating new observations within the existing time intervals. This is typically done to convert lower-frequency data into higher-frequency data. For example, converting daily data to hourly data.
    - To upsample the data, interpolation techniques such as linear interpolation, spline interpolation, or nearest-neighbor interpolation can be used to estimate the values for the new time points.
    - Upsampling can also involve introducing missing values or NaN (Not-a-Number) values for the new time points if there is no available data.
2. Downsampling (Decreasing Frequency):
    - Downsampling involves decreasing the frequency of the data by aggregating or summarizing the existing observations over larger time intervals. This is typically done to convert higher-frequency data into lower-frequency data. For example, converting hourly data to daily data.
    - Various aggregation methods can be used for downsampling, including taking the average, sum, maximum, minimum, or other statistical measures of the original observations within each new time interval.
    - Downsampling can also involve selecting a representative value from the original data, such as selecting the first value, the last value, or the value at a specific timestamp within each new time interval.

When resampling, it's important to consider the characteristics of the data and the objectives of the analysis. Some additional points to keep in mind:

- When upsampling, the interpolation method chosen should be appropriate for the data and the intended analysis. Linear interpolation is commonly used, but other methods can be more suitable for specific situations.
- When downsampling, the aggregation method selected should preserve the desired information from the original data. Different aggregation methods may be more appropriate depending on the nature of the variable and the analysis being performed.
- It's crucial to validate and assess the quality of the resampled data, especially when introducing new values or aggregating data. Ensure that the resampled data accurately represents the underlying patterns and trends in the original time series.

Resampling is a powerful technique for aligning, transforming, or summarizing time series data to meet the specific requirements of an analysis or model. The choice of resampling method depends on the nature of the data, the desired time scale, and the objectives of the analysis.

5. Smoothing: Smoothing techniques in time series analysis are methods used to remove noise or irregularities from a time series dataset in order to identify underlying patterns and trends more easily. These techniques help in reducing the effects of short-term fluctuations and random variations, allowing for a clearer representation of the underlying signal. Some commonly used smoothing techniques are as follows:
   (a) Moving Average: In this technique, a sliding window of fixed width moves across the time series, and the average of the data points within the window is calculated. This average value is then used as the smoothed value for that time point. Moving averages can be simple, where each data point is given equal weight, or weighted, where more recent data points are assigned higher weights.
   (b) Exponential Smoothing: This method assigns exponentially decreasing weights to past observations. It uses a weighted average approach, with more weightage given to recent observations and diminishing weightage as you move back in time. Exponential smoothing is widely used for forecasting and has different variations such as simple exponential smoothing, Holt's linear method, and Holt-Winters' method for seasonal data.
   (c) Savitzky-Golay Filter: This technique is commonly used for smoothing time series data, particularly when dealing with noisy data. It applies a polynomial regression within a moving window and replaces each data point with the value obtained from the regression. The Savitzky-Golay filter preserves the shape and trends of the original data while removing high-frequency noise.
   (d) LOESS (Locally Weighted Scatterplot Smoothing): LOESS is a non-parametric regression technique that estimates the relationship between variables based on local subsets of the data. It fits a separate regression line to different segments of the data and produces a smoothed curve that captures the underlying pattern without assuming a specific functional form.
   (e) Fourier Transforms: Fourier transforms are used to decompose a time series into its frequency components. Smoothing can be achieved by filtering out high-frequency noise or removing specific frequency components that are not of interest. Fourier smoothing techniques are particularly useful when dealing with periodic or seasonal data.
   (f) Kalman Filtering: Kalman filtering is an optimal recursive algorithm used to estimate the state of a dynamic system based on noisy observations. It is commonly used for smoothing and forecasting time series data, especially when the underlying system has a known linear structure.

The choice of technique depends on the characteristics of the data and the specific goals of the analysis. It is often helpful to experiment with different methods and compare the results to determine the most suitable approach for a given application.

6. Feature Extraction: Time series data can often be represented by a large number of data points, making it challenging to analyze. Feature extraction involves extracting relevant characteristics from the time series. Feature extraction in

time series analysis refers to the process of deriving relevant and informative characteristics, known as features, such as statistical measures (mean, variance), frequency domain features (FFT coefficients), or time-domain features (autocorrelation) from raw time series data. These features capture important patterns, trends, or statistical properties of the data and are used as input variables in various analysis or modeling tasks. These features can help in subsequent analysis or modeling.

Feature extraction helps simplify the data representation, reduce dimensionality, and enhance the performance of subsequent algorithms or models. Here are some common techniques used for feature extraction in time series analysis:

1. Statistical Features:
   - Mean: Average value of the time series.
   - Variance: Measure of the spread or dispersion of the data.
   - Skewness: Measure of the asymmetry of the data distribution.
   - Kurtosis: Measure of the peakedness or flatness of the data distribution.
   - Autocorrelation: Measure of the similarity between observations at different lags.
   - Percentiles: Values that divide the data into specific proportions (e.g., quartiles).
2. Frequency Domain Features:
   - Fast Fourier Transform (FFT) Coefficients: Magnitudes or phases of the frequency components in the time series.
   - Power Spectral Density: Distribution of power across different frequencies.
   - Wavelet Transform Coefficients: Representations of the time series at different scales and resolutions.
3. Time-Domain Features:
   - Rolling Statistics: Moving averages, standard deviations, or other statistical measures computed over a sliding window.
   - Lagged Values: Previous observations at specific lags.
   - Change-based Features: Differences or rates of change between consecutive observations.
   - Entropy: Measure of the unpredictability or complexity of the data.
4. Shape-based Features:
   - Slope: Trend or rate of change over a specific time interval.
   - Peaks and Valleys: Identifying the maximum and minimum points in the time series.
   - Shapelets: Subsequences or patterns that represent specific shapes or motifs in the time series.

5. Waveform Characteristics:
    - Rise and Fall Times: Duration of the ascending and descending parts of the waveform.
    - Amplitude: Maximum and minimum values within the waveform.
    - Waveform Moments: Statistical measures such as mean, variance, skewness, and kurtosis computed over the waveform.
6. Recurrence Plot-based Features:
    - Recurrence Quantification Analysis (RQA): Measures derived from recurrence plots, such as determinism, entropy, or laminarity.
    - Distance Measures: Distances between recurrence points or line segments in the recurrence plot.

These are just a few examples of feature extraction techniques used in time series analysis. The choice of features depends on the specific characteristics of the data, the analysis objectives, and the algorithms or models being used. It is often beneficial to combine multiple features to capture different aspects of the time series and to experiment with feature selection or dimensionality reduction methods to improve the efficiency and performance of the analysis.

7. Modeling and Analysis: Once the data is processed and transformed, various time series analysis techniques can be applied. This may include methods like autoregressive integrated moving average (ARIMA), seasonal decomposition of time series (STL), or more advanced approaches like state space models or deep learning models.

Modeling and analysis in time series refers to the process of developing mathematical or statistical models to understand and predict the behavior of the time series data. It involves applying various techniques and algorithms to uncover patterns, trends, relationships, and dependencies within the data, as well as making forecasts or extrapolations into the future. Here are some key steps involved in modeling and analyzing time series data:

1. Data Preprocessing: Before modeling, it's crucial to preprocess the time series data. This may include handling missing values, addressing outliers, normalizing or transforming the data, and ensuring the data is stationary if required.
2. Model Selection: Choose an appropriate model based on the characteristics of the time series data and the objectives of the analysis. Common time series models include autoregressive integrated moving average (ARIMA), exponential smoothing models, state space models, and recurrent neural networks (RNNs).
3. Model Fitting: Estimate the parameters of the chosen model using the available time series data. This involves optimization techniques such as maximum likelihood estimation (MLE) or least squares estimation (LSE) to find the best-fitting parameters that minimize the difference between the model and the observed data.

4. Model Evaluation: Assess the performance and goodness-of-fit of the model using appropriate evaluation metrics. This may include measures such as mean squared error (MSE), mean absolute error (MAE), Akaike Information Criterion (AIC), or Bayesian Information Criterion (BIC). Comparison with benchmark models or baseline models is also important.
5. Model Diagnostics: Examine the residuals or errors of the fitted model to check for any patterns, autocorrelation, heteroscedasticity, or other violations of assumptions. Diagnostic tests such as Ljung-Box test or Durbin-Watson test can be used to assess the quality of the model.
6. Forecasting and Prediction: Utilize the fitted model to make future predictions or forecasts. This involves projecting the time series values beyond the observed data and estimating the associated uncertainty or prediction intervals.
7. Model Interpretation and Analysis: Interpret the parameters and results of the model to gain insights into the underlying patterns, trends, and relationships within the time series. Analyze the coefficients or weights of the model to understand the contribution and significance of different factors or variables.
8. Sensitivity Analysis and Scenario Testing: Perform sensitivity analysis by assessing the impact of changes in model parameters or assumptions on the forecasts. Test different scenarios or what-if analyses to understand the potential outcomes under varying conditions.
9. Visualization and Reporting: Present the results of the modeling and analysis in visual and interpretable forms. Use plots, charts, and graphs to illustrate the patterns, trends, and forecasted values. Prepare a comprehensive report summarizing the methodology, findings, and conclusions.

The modeling and analysis process may involve iterations and refinements based on the results and insights gained. It's important to select appropriate models, validate their assumptions, and continually assess the accuracy and reliability of the forecasts to ensure robust and meaningful analysis of time series data.

8. Evaluation and Visualization: Finally, the results of the analysis need to be evaluated to assess the model's performance and its ability to capture the patterns and trends in the data. Visualizations such as line plots, scatter plots, or autocorrelation plots can be used to visualize the processed time series and the results of the analysis. Evaluation and visualization are crucial steps in time series analysis to assess the performance of models, validate assumptions, and effectively communicate the results. Here are some key aspects of evaluation and visualization in time series analysis:

    1. Evaluation Metrics:

        - Mean Squared Error (MSE): Measures the average squared difference between the predicted and actual values.
        - Mean Absolute Error (MAE): Measures the average absolute difference between the predicted and actual values.
        - Root Mean Squared Error (RMSE): Square root of the MSE, providing a measure in the same unit as the data.

- Mean Absolute Percentage Error (MAPE): Measures the average percentage difference between the predicted and actual values.
- Forecast Error Variance Decomposition (FEVD): Decomposes the forecast error variance into contributions from different factors or variables.

2. Accuracy and Residual Analysis:
   - Plotting Actual vs. Predicted: Visualize the actual and predicted values to assess the accuracy of the model.
   - Residual Analysis: Plot the residuals (errors) to check for patterns, autocorrelation, heteroscedasticity, or other violations of assumptions. Use diagnostic tests like Ljung-Box test or Durbin-Watson test.
3. Forecast Visualization:
   - Time Series Plots: Display the observed data and forecasted values on a time series plot to compare their trends and patterns.
   - Prediction Intervals: Visualize the prediction intervals to convey the uncertainty associated with the forecasts. This can be done using shaded regions or error bars around the point forecasts.
   - Rolling Forecast Origin: Plot the rolling forecasts over time to track the model's performance as new data becomes available.
4. Seasonality and Trend Analysis:
   - Seasonal Decomposition: Decompose the time series into its seasonal, trend, and residual components using methods like Seasonal Decomposition of Time Series (STL) or X-12-ARIMA.
   - Seasonal Subseries Plots: Plot subsets of the data corresponding to each season to observe any seasonal patterns or variations.
   - Trend Analysis: Plot the trend component of the decomposed time series to visualize the long-term trend.
5. Correlation and Autocorrelation Analysis:
   - Autocorrelation Function (ACF) Plot: Visualize the autocorrelation coefficients at different lags to identify significant lags and assess the presence of seasonality or dependence in the time series.
   - Partial Autocorrelation Function (PACF) Plot: Examine the partial autocorrelation coefficients to identify the order of autoregressive (AR) terms in the model.
6. Heatmaps and Contour Plots:
   - Heatmaps: Use color-coded heatmaps to display patterns in multivariate time series data or correlation matrices.
   - Contour Plots: Plot two variables against time on a 2D contour plot to visualize the joint behavior and relationships over time.

7. Interactive Visualizations:
    - Interactive Dashboards: Build interactive dashboards or applications to allow users to explore and interact with time series data, select variables, adjust parameters, and visualize results dynamically.
    - Interactive Plots: Use tools like Plotly or Bokeh to create interactive plots, allowing users to zoom, pan, and hover over data points for detailed information.

The choice of evaluation metrics and visualization techniques depends on the specific objectives, characteristics of the time series data, and the models used. Effective evaluation and visualization help in understanding the model's performance, identifying areas of improvement, validating assumptions, and effectively communicating the insights and findings derived from the analysis of time series data.

3.14 Graphical Representation of Time Series

There are several graphical representations commonly used to visualize time series data. These visualizations help to understand the patterns, trends, and other characteristics present in the data. Here are some of the most commonly used graphical representations of time series:

1. Time Plot: A time plot is a simple line graph where the time series observations are plotted on the y-axis against time on the x-axis. This plot shows the changes in the values of the time series over time, allowing for visual inspection of trends, seasonality, and other patterns.
2. Seasonal Plot: A seasonal plot is a variation of the time plot that focuses on the seasonal patterns in the data. It displays the data for multiple seasonal periods, typically in a single year, in separate panels or overlaid on top of each other. This plot helps to identify repeating patterns and seasonality effects.
3. Scatter Plot: A scatter plot is useful when examining the relationship between two variables in a time series. It displays individual data points as dots on a graph, with one variable on the x-axis and the other on the y-axis. This plot can reveal any correlation or patterns between the two variables.
4. Histogram: A histogram is a graphical representation that shows the distribution of values in a time series. It divides the range of values into bins and displays the frequency or count of observations falling into each bin. Histograms can provide insights into the shape of the distribution, such as whether it is symmetric, skewed, or has multiple peaks.

5. Box Plot: A box plot, also known as a box-and-whisker plot, provides a summary of the distribution of a time series. It displays the minimum, maximum, median, and quartiles of the data. This plot can help identify outliers, the spread of the data, and the presence of skewness.
6. Autocorrelation Plot: An autocorrelation plot, also called a correlogram, shows the correlation of a time series with its lagged values. It plots the autocorrelation coefficient on the y-axis against the lag on the x-axis. This plot is useful for identifying any significant lagged relationships or seasonality effects.
7. Spectral Plot: A spectral plot, often obtained through a technique called spectral analysis, provides insights into the frequency components present in a time series. It displays the power or amplitude of each frequency on the y-axis against frequency on the x-axis. This plot can be useful for detecting periodicities and dominant frequencies in the data.

These graphical representations can be created using various software tools like Python's Matplotlib, R's ggplot2, or other specialized time series analysis software. Each plot provides a different perspective on the time series data and can aid in identifying patterns, trends, seasonality, outliers, and other important features of the data.

3.15 Time Series Visualization: Techniques and Examples

Time series visualization plays a crucial role in understanding patterns, trends, and anomalies in temporal data. By employing effective visualization techniques, we can gain valuable insights and make informed decisions. This article explores various visualization methods for time series data, accompanied by examples that demonstrate their application.

1. Line Plots: Line plots are fundamental for visualizing time series data. They show the trend and fluctuations over time, providing a clear depiction of how values change. For instance, a line plot can be used to visualize the daily closing prices of a stock over a year, revealing any upward or downward trends.
2. Seasonal Plots: Seasonal plots help identify recurring patterns within a time series. By aggregating data based on specific time intervals, such as months or quarters, and plotting them together, seasonal patterns become evident. An example would be a seasonal plot showing the monthly average temperature variations throughout the year.
3. Decomposition Plots: Decomposition plots help decompose a time series into its constituent components: trend, seasonality, and residuals. This allows for a detailed examination of each component's contribution to the overall behavior of the series. For instance, decomposing a retail sales time series might reveal an increasing trend, seasonal spikes during holidays, and random fluctuations due to external factors.

4. Lag Plots: Lag plots, also known as scatterplots, help visualize the correlation between a time series and its lagged values. By plotting a time series against its lagged version, we can assess if there is any autocorrelation present. For example, a lag plot might show a positive correlation between today's stock prices and prices from the previous day, indicating some degree of persistence.
5. Boxplots: Boxplots provide a concise summary of the distribution of a time series at different time points or groups. By displaying quartiles, outliers, and median values, boxplots help identify variations, central tendencies, and potential anomalies. Boxplots can be useful in visualizing monthly sales for multiple years, allowing for comparisons and the detection of outliers.
6. Heatmaps: Heatmaps are effective for displaying time series data with multiple dimensions or variables. They use colors to represent values, allowing patterns and relationships to be easily identified. For instance, a heatmap can visualize hourly electricity consumption over several months, highlighting peak and off-peak hours across weekdays and weekends.
7. Interactive Visualizations: Interactive visualizations, such as interactive line plots or interactive heatmaps, enable users to explore time series data dynamically. They offer features like zooming, panning, and tooltips, enhancing the ability to analyze and interact with the data. Interactive visualizations can be particularly useful when dealing with large datasets or when examining detailed patterns.

Time series visualization techniques play a vital role in understanding and interpreting temporal data. Line plots, seasonal plots, decomposition plots, lag plots, boxplots, heatmaps, and interactive visualizations are powerful tools that aid in uncovering patterns, trends, and anomalies. By utilizing these techniques, analysts can gain valuable insights and make informed decisions based on the visual exploration of time series data.

Remember, in practice, the choice of visualization techniques depends on the specific characteristics of the time series data and the insights sought. Experimenting with various visualization methods can lead to a deeper understanding of the underlying patterns and facilitate effective communication of findings.

Note: Time series graphs and visualization are closely related but not exactly the same.

A time series graph specifically refers to a graphical representation of the time series data, where the observations are plotted against time. It is a specific type of visualization that focuses on displaying the temporal patterns and trends in the data.

Time series graphs commonly use a line plot, where the time points are represented on the x-axis, and the corresponding values of the time series are plotted on the y-axis. This type of graph allows for the visual examination of the changes and fluctuations in the data over time. Time series graphs may also include additional elements such as trend lines, seasonal components, or confidence intervals to enhance the understanding of the data.

On the other hand, data visualization is a broader term that encompasses various techniques and approaches to represent data visually. It includes not only time series

graphs but also other types of visualizations such as scatter plots, histograms, bar charts, heatmaps, and more.

Data visualization aims to present data in a visually appealing and informative way, allowing users to explore patterns, relationships, and insights that might not be apparent in raw data. While time series graphs are a specific type of data visualization focused on time-dependent data, there are numerous other visualization techniques that can be applied to time series or other types of data.

In summary, time series graphs are a specific type of visualization technique used to represent time series data, while data visualization encompasses a broader range of techniques used to represent data in visual form. Time series graphs are a subset of data visualization techniques that specifically focus on representing time-dependent patterns and trends.

3.16 Additional Topics

Autocorrelation and partial autocorrelation are important concepts in time series analysis that help identify and model the temporal dependencies or relationships within a time series. They provide insights into the lagged relationships between the observations and are useful for determining the appropriate order of autoregressive (AR) and moving average (MA) components in time series models. Let's delve into the definitions and applications of autocorrelation and partial autocorrelation:

Autocorrelation (ACF): Autocorrelation measures the correlation between a time series and its lagged values. It quantifies the linear relationship between a data point and its historical observations at different time lags. The autocorrelation function (ACF) is commonly used to plot and analyze the autocorrelation.

The ACF at lag k, denoted as ACF(k), measures the correlation between the time series at time t and the time series at time t-k. A positive autocorrelation indicates a positive linear relationship between the current observation and the observation at the lagged time point, while a negative autocorrelation indicates a negative linear relationship.

The ACF plot visualizes the autocorrelation at different lags. It helps identify significant lagged relationships and patterns in the time series. For example, if the ACF plot shows significant autocorrelation at lag 1 and gradually diminishes as the lag increases, it suggests the presence of an autoregressive (AR) component in the time series.

Partial Autocorrelation (PACF): The partial autocorrelation measures the linear relationship between two variables while controlling for the influence of the intermediate observations. It quantifies the direct association between a data point and its historical observations, removing the indirect effects through the intermediate lags.

The partial autocorrelation function (PACF) is used to plot and analyze the partial autocorrelation. The PACF at lag k, denoted as PACF(k), measures the correlation between the time series at time t and the time series at time t-k, considering the intermediate lags.

The PACF plot displays the partial autocorrelation at different lags. It helps identify the significant direct relationships between observations and is particularly useful in determining the order of autoregressive (AR) components in a time series model. In the PACF plot, significant partial autocorrelation at lag k suggests the inclusion of an AR(k) component in the model.

By examining the ACF and PACF plots and identifying significant autocorrelation and partial autocorrelation, we can determine the appropriate order of AR and MA components in autoregressive integrated moving average (ARIMA) models or other time series models.

These tools enable us to model and understand the temporal dependencies and patterns in time series data, leading to more accurate modeling, forecasting, and analysis of time-dependent phenomena.

ARIMA models are widely used in time series analysis for modeling and forecasting data. ARIMA models combine autoregressive (AR), moving average (MA), and differencing components to capture the temporal dependencies and patterns present in a time series. The acronym ARIMA stands for:

1. Autoregressive (AR) Component: The autoregressive component captures the linear relationship between the current value of the time series and its past values. It assumes that the current value is a linear combination of its lagged values and a white noise error term. The order of the AR component is denoted by the parameter p, representing the number of lagged terms included in the model.
2. Integrated (I) Component: The integrated component refers to differencing the time series to make it stationary. Differencing helps remove trends and non-stationarity from the data. The differencing order is denoted by the parameter d, representing the number of times differencing is applied to achieve stationarity.
3. Moving Average (MA) Component: The moving average component captures the linear relationship between the current value of the time series and past error terms (residuals). It assumes that the current value is a linear combination of the error terms and a white noise error term. The order of the MA component is denoted by the parameter q, representing the number of lagged error terms included in the model.

ARIMA models are specified as ARIMA(p, d, q), where p, d, and q represent the order of the AR, I, and MA components, respectively.

The steps involved in building an ARIMA model are as follows:

1. Data Preparation: Preprocess the time series data by handling missing values, outliers, and transforming it to achieve stationarity if necessary.
2. Model Identification: Determine the appropriate orders (p, d, q) by analyzing the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots of the differenced data. These plots help identify the significant lagged relationships and guide the selection of the model orders.
3. Parameter Estimation: Use maximum likelihood estimation or other optimization techniques to estimate the model parameters. This involves fitting the AR, I, and MA components to the data.

4. Model Diagnostic Checking: Evaluate the residuals of the model to ensure that they follow the assumptions of white noise (i.e., no systematic patterns or correlations). Diagnostic checks include examining the autocorrelation of residuals and performing statistical tests.
5. Model Forecasting: Use the estimated model parameters to forecast future values of the time series. Forecasting can be done using various techniques, such as one-step ahead forecasting or multistep ahead forecasting.

ARIMA models provide a flexible framework for modeling a wide range of time series data, capturing various temporal patterns and dependencies. ARIMA models are widely applicable in numerous domains, including finance, economics, sales forecasting, demand planning, supply chain management, environmental sciences, and many others. Their flexibility, interpretability, and ability to capture both short-term and long-term dependencies make them valuable tools for analyzing and understanding time-dependent data.

ARIMA models are versatile and widely used in time series analysis. They offer several applications and benefits in understanding, modeling, and forecasting time-dependent data. Here are some common usages of ARIMA models:

1. Trend Analysis and Forecasting: ARIMA models are effective for analyzing and forecasting time series data with trend components. They capture the linear relationships between the current value and its lagged values, allowing for trend estimation and prediction. ARIMA models can identify and model trends in economic indicators, stock prices, population growth, and other time-dependent phenomena.
2. Seasonal Analysis and Forecasting: Seasonal ARIMA (SARIMA) models, which extend the basic ARIMA framework to incorporate seasonality, are used to analyze and forecast data with seasonal patterns. SARIMA models capture both the autoregressive and moving average components along with seasonal differencing. They are valuable for understanding and predicting seasonal behavior in areas such as sales forecasting, demand planning, and climate analysis.
3. Anomaly Detection: ARIMA models can be employed for detecting anomalies or outliers in time series data. By fitting the model to the historical data and comparing the observed values with the model's predicted values, deviations from the expected behavior can be identified as potential anomalies. This is useful for detecting unusual events, outliers, or data points that deviate significantly from the normal pattern.
4. Time Series Decomposition: ARIMA models are used for decomposing a time series into its underlying components, such as trend, seasonality, and residuals. This decomposition helps isolate and analyze the individual elements of the time series, providing insights into their respective contributions and behaviors.
5. Forecasting Future Values: ARIMA models are primarily utilized for time series forecasting. By estimating the model parameters and utilizing past observations, ARIMA models can generate predictions for future values. The forecasted values provide valuable insights for decision-making, planning, resource allocation, and operational management in various domains.

6. Impact Analysis and Scenario Planning: ARIMA models can be employed to assess the potential impact of specific events or interventions on a time series. By modifying the model inputs or parameters, it is possible to simulate different scenarios and predict the resulting effects on the time series. This aids in evaluating the potential outcomes of various interventions or policy changes.
7. Model Selection and Comparison: ARIMA models serve as a benchmark or reference when comparing and evaluating the performance of other time series models. By establishing the predictive accuracy and goodness of fit of an ARIMA model, it becomes easier to assess the effectiveness and relevance of alternative modeling techniques.

ARCH (Autoregressive Conditional Heteroscedasticity) and GARCH (Generalized Autoregressive Conditional Heteroscedasticity) models are widely used in time series analysis to model and forecast volatility, particularly in financial data. These models are designed to capture the time-varying variance, or heteroscedasticity, often observed in financial time series.

ARCH Models: ARCH models were introduced by Engle (1982) and are based on the idea that the variance of a time series is autocorrelated and depends on past squared errors or residuals. The key assumption of ARCH models is that the conditional variance is a function of lagged squared residuals.

In an ARCH(p) model, the current conditional variance is modeled as a linear combination of the past p squared residuals, where p represents the order of the ARCH model. The ARCH(p) model is given by:

$$\text{Var}(t) = \alpha_o + \alpha_1 \varepsilon^2(t-1) + \alpha_2 \varepsilon^2(t-2) + \cdots + \alpha_p \varepsilon^2(t-p)$$

Here, Var(t) represents the conditional variance at time t, ε(t) is the residual at time t, and $\alpha_O, \alpha_2, \ldots, \alpha_p$ are the model parameters that need to be estimated. The ARCH(p) model captures the autocorrelation in squared residuals, allowing for the modeling of time-varying volatility.

GARCH Models: GARCH models, introduced by Bollerslev (1986), are an extension of ARCH models that incorporate both autoregressive and moving average components to capture the volatility dynamics more accurately.

In a GARCH(p, q) model, the conditional variance at time t is modeled as a linear combination of past squared residuals and past conditional variances. The GARCH(p, q) model is given by:

$$\begin{aligned}\text{Var}(t) &= \alpha_o + \alpha_1 \varepsilon^2(t-1) + \alpha_2 \varepsilon^2(t-2) + \cdots + \alpha_p \varepsilon^2(t-p) \\ &+ \beta_1 \text{Var}(t-1) + \beta_2 \text{Var}(t-2) + \cdots + \beta_q \text{Var}(t-q)\end{aligned}$$

The additional terms $\beta_1 \text{Var}(t-1) + \beta_3 \text{Var}(t-2) + \cdots + \beta_q \text{Var}(t-q)$ capture the autoregressive behavior of the conditional variance itself. The GARCH(p, q) model captures both the short-term and long-term persistence of volatility.

Estimation and Inference: The parameters of ARCH and GARCH models can be estimated using various methods, such as maximum likelihood estimation or

the method of moments. Estimation involves optimizing the likelihood function or minimizing the sum of squared residuals.

Model diagnostic checks are crucial to ensure that the model adequately captures the volatility dynamics. These checks include examining the residuals for autocorrelation, normality, and other statistical properties. If the residuals exhibit residual autocorrelation, other model specifications or extensions may be considered.

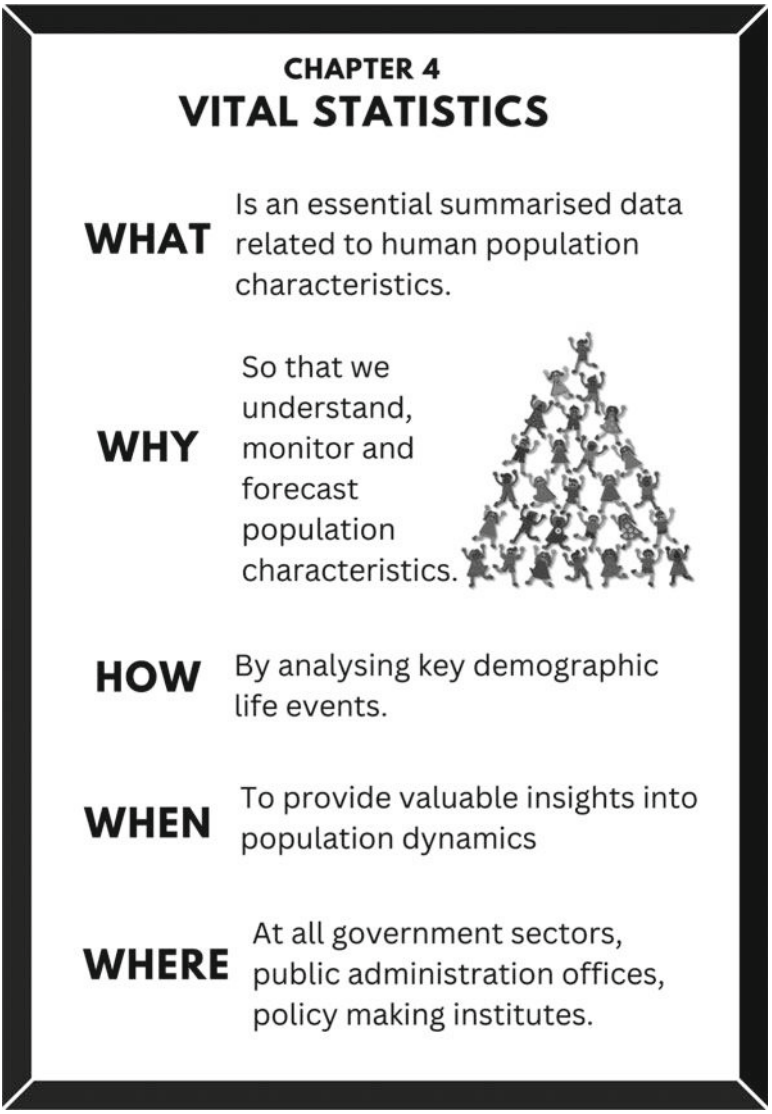
Applications: ARCH/GARCH models find extensive applications in financial econometrics, particularly in modeling and forecasting asset returns, volatility, and risk. They provide valuable insights into the time-varying volatility patterns, conditional variances, and risk measures.

Some common applications of ARCH/GARCH models include:

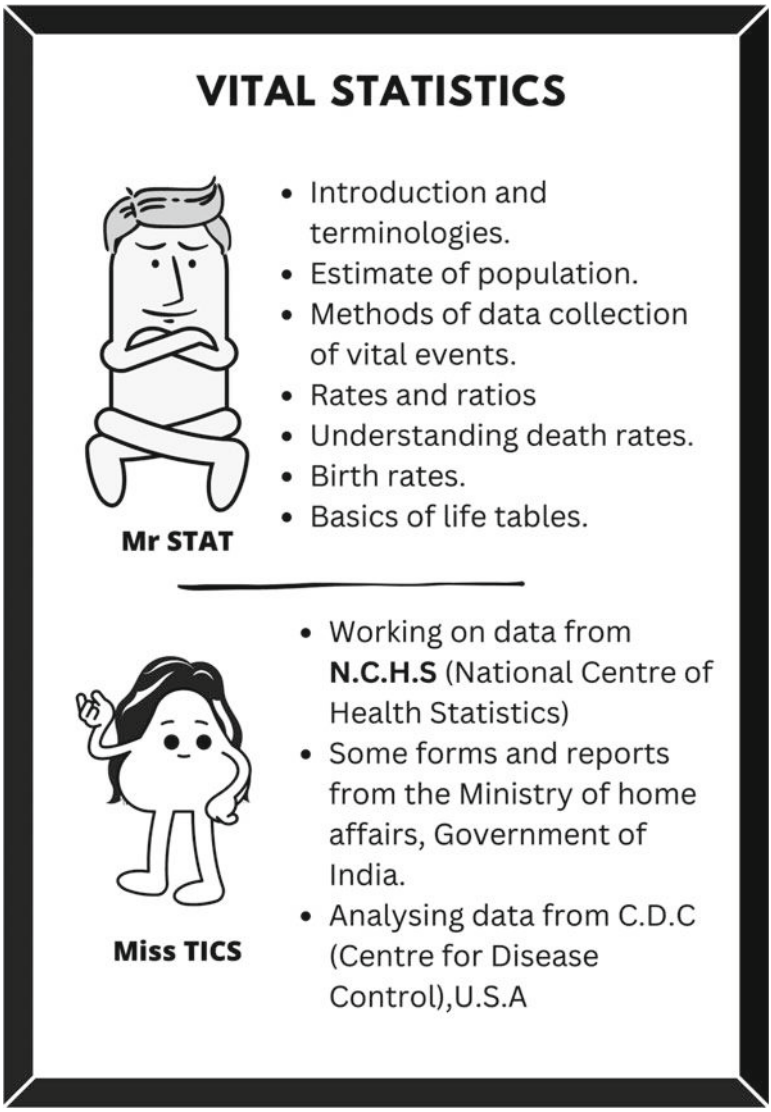
1. Volatility Forecasting: ARCH/GARCH models allow for accurate and dynamic forecasting of volatility, which is essential for risk management, option pricing, and portfolio optimization.
2. Risk Measurement: ARCH/GARCH models provide measures of conditional variances and volatility that are crucial for estimating risk measures such as Value at Risk (VaR) and Expected Shortfall (ES).
3. Portfolio Optimization: By incorporating volatility forecasts from ARCH/GARCH models, portfolio managers can better estimate risk and construct optimal portfolios.
4. Financial Market Analysis: ARCH/GARCH models help analyze the behavior of financial time series, investigate volatility clustering, and examine the impact of news and events on market volatility.

In summary, ARCH and GARCH models have become prominent tools in modeling and forecasting volatility in financial time series. By capturing time-varying variances and autocorrelation in squared residuals, these models provide valuable insights into the risk dynamics of financial assets, enabling better risk management and decision-making.

Chapter 4
Vital Statistics



S. Prasad, *Advanced Statistical Methods*, https://doi.org/10.1007/978-981-99-7257-9_4



What are vital statistics?

Vital statistics refer to quantitative data about vital events in a population. These events include births, deaths, marriages, divorces, and adoptions. Vital statistics provide details about these events, including dates, locations, and characteristics of the individuals involved.

How are vital statistics collected?

Vital statistics are typically collected by government agencies responsible for civil registration systems. These systems record and document vital events that occur within a jurisdiction. Data is collected through the completion of legal documents such as birth certificates, death certificates, marriage licenses, and divorce decrees. These documents are filled out by individuals, healthcare professionals, or other authorized personnel and are then registered with the relevant government agency.

When are vital statistics used?

Vital statistics are used for various purposes, including:

- Demographic analysis: Vital statistics help in studying and understanding population dynamics, such as birth rates, death rates, and migration patterns. These statistics provide insights into population growth, aging trends, and changes in the population's structure.
- Public health monitoring: Vital statistics help public health officials monitor and assess the health of a population. They can track mortality rates, causes of death, and disease prevalence, enabling the identification of health trends and the development of targeted interventions.
- Social and economic planning: Vital statistics are crucial for social and economic planning at both national and local levels. They inform policymakers and researchers about population trends, fertility rates, life expectancy, and other factors that influence resource allocation, infrastructure development, and social programs.
- Research and policy development: Vital statistics serve as a foundation for scientific research and policy development in various fields, including sociology, demography, public health, and economics. Researchers and policymakers rely on these statistics to inform their work, make evidence-based decisions, and develop effective strategies.

Why are vital statistics important?

Vital statistics play a vital role in society for several reasons:

- Policy formulation and evaluation: Governments use vital statistics to develop and evaluate policies related to healthcare, education, social welfare, and other areas. Accurate and up-to-date data on vital events help policymakers make informed decisions and assess the impact of their policies.
- Resource allocation: Vital statistics guide the allocation of resources, such as healthcare facilities, educational institutions, and social services. Understanding population dynamics helps in identifying areas with specific needs and distributing resources accordingly.
- Monitoring progress: Vital statistics enable the monitoring of progress toward national and international goals. For example, they help track progress in achieving targets related to reducing child mortality, improving maternal health,

or combating communicable diseases, as set by the United Nations' Sustainable Development Goals (SDGs).
- Historical and genealogical research: Vital statistics are valuable for historical and genealogical research, allowing individuals to trace their ancestry and understand their family history. These records provide insights into past populations, migration patterns, and societal changes.

How do vital statistics work?

- Data Collection: Vital statistics data is collected from various sources, including civil registration systems, hospitals, health departments, and census surveys. Governments and relevant organizations typically manage the collection and compilation of this data.
- Registration of Vital Events: Vital events like births, deaths, marriages, and divorces are officially registered by the concerned authorities. This registration ensures that accurate and comprehensive data is available for analysis.
- Data Processing: Once collected, the data is processed, verified, and compiled into databases or statistical systems. The data is organized and made accessible for analysis and research.
- Data Analysis: Researchers and statisticians analyze the data to derive meaningful insights. They study trends, patterns, and relationships between various vital events to understand the population's characteristics and changes over time.
- Demographic Indicators: Vital statistics are used to calculate various demographic indicators, such as birth rates, death rates, fertility rates, infant mortality rates, life expectancy, and migration rates. These indicators provide valuable information about population health, age distribution, and other demographic factors.
- Policy and Planning: Governments, policymakers, and public health officials use vital statistics to make informed decisions about resource allocation, healthcare planning, and social policy. The data helps identify areas of concern and formulate strategies to address specific demographic and health challenges.
- Public Health Surveillance: Vital statistics play a crucial role in public health surveillance. Monitoring mortality rates, disease outbreaks, and other health indicators helps identify emerging health threats and assess the effectiveness of public health interventions.

4.1 Introduction

This is a branch of statistics that accounts for every vital event in human life in a legal way for harmony in society. Vital statistics are a fundamental component of demographic and public health research, providing essential information about the population's characteristics, events, and changes over time. These statistics encompass a range of crucial data, such as births, deaths, marriages, annulments, separations, adoptions and divorces, forming the backbone of demographic analysis and health surveillance. By meticulously recording and analyzing vital events, governments

and organizations gain valuable insights into population dynamics, health trends, and social patterns. These statistics play a pivotal role in guiding policy decisions, resource allocation, and the development of targeted interventions to promote the well-being of societies worldwide. In this context, the accurate and comprehensive collection of vital statistics is vital to understanding and addressing the needs and challenges of diverse communities in a rapidly evolving world (Fig. 4.1).

The term vital statistics is also used for individual measures of these vital events. Thus, a birth rate is an example of vital statistics and analysis of birth rate trends is an example of a vital statistic application. Other demographically significant life events such as change of residence (migration), change of citizenship (naturalization), and name changes are not recorded, mainly because information on these is usually obtained from other statistical systems such as population registers. They are an important national source of information for understanding public health (Table 4.1).

Every country will have a division for collecting and maintaining the records of vital events of people under the National Centre for Health Statistics (NCHS). The categories of data collected are:

Fig. 4.1 Vital statistics—a study of population demographics

Table 4.1 Details collected and their respective statistical units mentioned in the public records

	DETAILS COLLECTED	STATISTICAL UNITS
	• Number of people born. • Gender. • Health condition.	• Birth rates. • Infant mortality rate
	• Education attained. • Institution of Study. • Job and nature of work.	• Literacy rate. • Rate of unemployment. • Rate of child labor.
	• Health status. • Emotional and social well being. • Cost of good health.	• Poverty rates. • Public health conditions. • Rate of migrations/ emmigrations.
	• Wedding. • Age of marriage. • Place of stay.	• Rates of divorce cases.
	• Death • Age at the time of death • Reason for death.	• Death rates. • Rate of death due to cancer. • Age Specific Death Rates.

- Birth data.
- Mortality data.
- Fetal death data.
- Marriages and divorces.
- Survey reports on maternal and infant health.
- Survey reports of mortality follow-back survey.

4.2 Advantages of Vital Statistics

For an individual**:**

- Vital statistics are much for much use of an individual and the family.
- Birth certificate issued by the registering authority is a crucial document that identifies the existence and identity of a child in society.
- Marital status of an individual is recognized and acknowledged with the formal marriage certificate issued by the registrar in a form of acceptance from the government for two individuals to live together.

For legal aspects:

- Registering for all the vital events is mandatory. Government records are valid proof of the land owned and other valuable assets in your name, your relationship status, and your sole existence in society.
- This will keep the individual safe from fraudsters.

For administrators/planning committees.

- Investments and allocations in the budget for various programs that promote societal health consider the vital statistical data as a base.
- Population trends analyzed thoroughly for city expansion projects, population control programs, etc. Hospitals and medical facilities are provided based on the health/economic conditions of society.
- Vital statistics serve as the foundation for actuarial science, which includes life insurance.
- Policy planners, administrators, and others need vital statistics to estimate population trends and forecasts, and these are required for planning and evaluation of economic and social development programs.
- Death rates are used by epidemiologists and medical researchers to identify hazardous occupations, compute the life expectancy of an individual in certain areas, etc.

For the world and its health:

- UNESCO, UNICEF, WHO, and ILO are few organizations that served the world, and a general understanding of trends and patterns in the population of the specific region and specific period under study is necessary before any welfare schemes and programs are compiled.
- The mortality figures aid in our efforts to enhance community health. For instance, statistics on communicable diseases assist the health authorities in enhancing the hygienic state of the affected area and medical institutions.
- These statistics help us to predict the future population structure of a country or region.

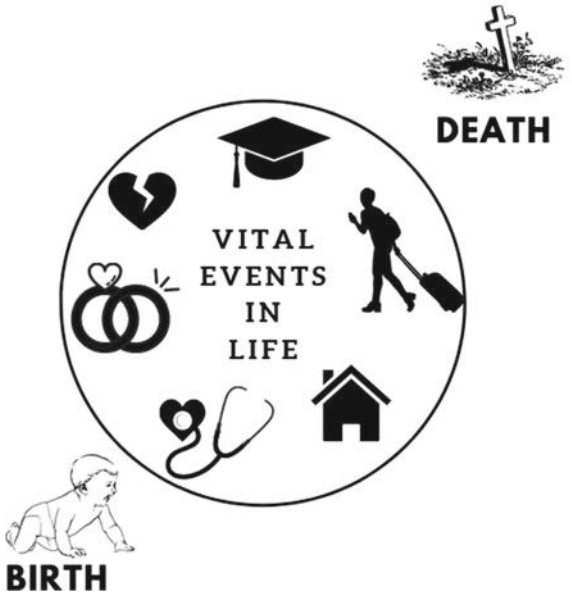


Fig. 4.2 All the important life events from birth to death of a human

- To get an idea about the changes in the population profile of a region, we need vital statistics in terms of age, gender, religion, births, deaths, migration, marriages, etc (Fig. 4.2).

4.3 Common Terminologies Used in Vital Statistics

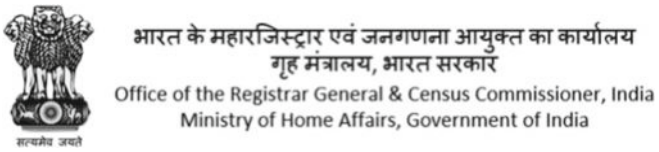
1. Live Births: A live birth is defined as the delivery of a child who displayed any kind of life; the term "live births" refers to births overall, excluding stillbirths.
2. Marriage: Acceptance of two people who agree to love together adhering to the societal and legal rules.
3. Death: An irreversible cessation of circulatory/respiratory functions or the entire brain and brain stem is considered as death.
4. Fatal death: Refers to the intrauterine death of a fetus at any time during pregnancy.
5. Still birth: This refers to the delivery of a viable fetus dead. They usually occur in the later pregnancy periods such as, post 20/28 weeks of pregnancy.

6. Vital events: All the important events in one's life starting from birth to death are termed as vital events, such as birth, marriage, divorce, adoption, partnership in business, purchase of land/asset, and death. All the vital events are registered in the civil register as a proof of record.
7. Cohort: Hypothetical group of people under study.
8. Morbidity: Refers to the state of having a disease or being exposed to a disease in the population. At times, morbidity refers to medical problems that may or may not arise due a treatment.

4.4 Sources of Data in Vital Statistics:

The three methods of data collection for the vital statistics system are:

1. Registration Method:



This method consists of the continuous and permanent recording of births, deaths, marriages, migrations, etc. This practice is made mandatory by the government, and the documents are well-verified and certified by the allotted departments.

The registration of birth and death is done under the provisions of a central Act namely the Registration of Births and Deaths (RBD) Act of 1969 and State Rules framed on the basis of Model Rules, 1999.

The report "Vital Statistics based on Civil Registration System" for the year 2019 at the national level has been released on June 15, 2021. The proportion of registered births and deaths has witnessed a steady increase over the years. The registration level of births for the country has gone up to 92.7% in 2019 from 82.4% in 2011, whereas on the other hand, registration level of deaths during 2019 has increased to 92.0% from 66.4% in 2011.

Here is a sample birth certificate document issued by the state of Maharashtra (Fig. 4.3).

2. Census Method:

This is the term used for enumerating the population of a country and provides the most important statistical information such as age, gender, marital status, level

प्रमाणपत्र क्रमांक/ Certificate No. ..... नमुना - ५ / Form - 5

महाराष्ट्र शासन
GOVERNMENT OF MAHARASHTRA
आरोग्य विभाग
HEALTH DEPARTMENT
प्रमाणपत्र निर्गमित करणाऱ्या स्थानिक क्षेत्राचे नाव
Name of local body issuing certificates

सत्यमेव जयते

जन्म प्रमाणपत्र
BIRTH CERTIFICATE

(जन्म व मृत्यु नोंदणी अधिनियम, १९६९ च्या कलम १२/१७ आणि महाराष्ट्र जन्म आणि मृत्यु नोंदणी नियम, २००० चे नियम ८/१३ अन्वये देण्यात आले आहे.)

(Issued under section 12/17 of the Registration of Births & Deaths Act, 1969 and Rule 8/13 of the Maharashtra Registration of Births and Deaths Rules, 2000.)

प्रमाणित करण्यात येत आहे की, खालील माहिती जन्माच्या मूळ अभिलेखाच्या नोंदवहीतून घेण्यात आली आहे, जी की (स्थानिक क्षेत्र) .....,तालुका ....., जिल्हा ....., महाराष्ट्र राज्या च्या नोंदवहीत उल्लेख आहे.

This is to certify that the following information has been taken from the original record of birth which is the register for (local area/local body) _____ of tahsil /block _____ of District _____ of **Maharashtra State.**

बाळाचे नाव :	लिंग : .....
Name of child : _____	Sex : _____
जन्म दिनांक :	जन्म ठिकाण : .....
Date of Birth : _____	Place of birth : _____
आईचे पूर्ण नाव :	वडिलांचे पूर्ण नाव : .....
Name of Mother : _____	Name of Father : _____
बाळाचे जन्माचे वेळी आई वडिलांचा पत्ता :	आई वडिलांचा कायमचा पत्ता : .....
.....	.....
.....	.....
Address of parents at the time of birth of the child :	Permanent address of Parents :
_____	_____
_____	_____
नोंदणी क्रमांक : _____	नोंदणी दिनांक : _____
Registration No. : _____	Date of Registration : _____
शेरा :	
Remarks (If any) : _____	निर्गमित करणाऱ्या प्राधिकाऱ्याची सही
	Signature of the issuing authority
प्रमाणपत्र दिल्याचा दिनांक :	प्राधिकाऱ्याचा पत्ता : .....
Date of issue : _____ शिक्का / Seal	Address of the issuing authority : _____

"प्रत्येक जन्म आणि मृत्यूची घटना नोंदल्याची खात्री करा" "Ensure Registration of every birth & death"

Fig. 4.3 Birth certificate issued by the government of Maharashtra, health department

of education, occupation, and religion. However, since it is conducted once in ten years, data for the other years is calculated using mathematical formulae. Population censuses typically use one of the two approaches:

- De facto—meaning enumeration of individuals as of where they are found in the census, regardless of where they normally reside.

- De jure—meaning enumeration of individuals as of where they usually reside, regardless of where they are on census day.

3. Survey—Along with census enumerations, there are various other surveys conducted for specific research and also with a requirement of collecting comprehensive data. Few of them are
   - National Sample Survey Organisation. (NSSO).
   - National Family Health Survey (NFHS).
   - District Level Household Surveys (DLHS-RCH).

4.5 Measurement of Population

The measurement of population in vital statistics involves the collection, recording, and analysis of data related to various vital events that occur within a population. Vital statistics primarily focus on four key components: births, deaths, marriages, and divorces. These events are crucial for understanding the dynamics of a population and providing essential information for demographic and public health research.

1. Births: The measurement of births involves recording the number of live births within a specified geographic area and time period. This data includes information about the child, such as date and time of birth, gender, birth weight, and the parents' demographic characteristics.
2. Deaths: Vital statistics capture data on deaths, providing insights into mortality patterns within a population. Information collected includes the date, time, and cause of death, as well as demographic details of the deceased individual.
3. Marriages: The measurement of marriages involves recording the number of legal unions between individuals within a specific jurisdiction and time frame. Data collected typically includes the date and location of the marriage and the demographic characteristics of the spouses.
4. Divorces: Vital statistics also encompass data on divorces, indicating the dissolution of marriages within a population. This information includes the date and location of the divorce and the demographic details of the individuals involved.

To ensure the accuracy and comprehensiveness of vital statistics, governments and relevant organizations establish registration systems. Civil registration systems, often managed by national or regional authorities, are responsible for recording and maintaining vital events' data. These systems serve as the primary source of vital statistics, and the data collected is usually compiled into vital registration reports.

The measurement of population through vital statistics is a vital aspect of demographic analysis, as it enables researchers, policymakers, and public health experts to gain insights into population trends, growth rates, mortality rates, age structures, and other demographic characteristics. This information is critical for making informed decisions, planning public services, and addressing the evolving needs of a population.

4.5.1 Calculation of Intercensal Estimates

Intercensal estimates refer to population estimates that are made for the time period between two consecutive decennial censuses. In many countries, official censuses are conducted every ten years to collect detailed demographic data about the entire population. However, the information obtained from a single census becomes outdated quickly due to population growth, migration, and other demographic changes. Intercensal estimates bridge the gap between two censuses, providing up-to-date and reliable population figures for each year.

Here are some key aspects of intercensal estimates:

1. Purpose: The primary purpose of intercensal estimates is to provide policymakers, government agencies, researchers, and the public with accurate and timely information about the current population size and its characteristics. These estimates help in planning and allocating resources for various services such as education, healthcare, housing, and infrastructure.
2. Data Sources: Intercensal estimates are typically based on a combination of data sources. These may include the most recent census data, administrative records (e.g., birth and death registrations, immigration records), and survey data (e.g., household surveys). Data from other relevant sources, such as tax records and social security data, may also be used to improve the accuracy of the estimates.
3. Methods: Various statistical methods are employed to develop intercensal estimates. One common approach is the cohort-component method, which projects population changes by age and sex using data on births, deaths, and net migration during the intercensal period. Other methods, such as time series analysis and demographic modeling, may also be used depending on data availability and the level of detail required.
4. Demographic Components: Intercensal estimates typically involve the analysis of three key demographic components: births, deaths, and migration. By tracking changes in these components, statisticians can estimate the population's growth or decline during the intercensal period.

5. Confidence Intervals: Intercensal estimates are not exact figures but rather statistical estimates. To account for uncertainty in the data and methods used, confidence intervals are often provided. These intervals represent a range within which the true population figure is likely to lie with a certain level of confidence.
6. Revisions: As more accurate data becomes available or methodologies are refined, intercensal estimates may be subject to revisions. Governments and statistical agencies usually release revised estimates periodically to reflect these improvements.

Intercensal estimates play a crucial role in maintaining updated population data between official censuses. They are essential tools for policymakers and researchers in understanding population dynamics, identifying demographic trends, and making informed decisions about public policy and resource allocation.

Methods for Calculating Intercensal Data

1. Linear Interpolation Method

$$\mathbf{P_t} = \mathbf{P}_0 + \frac{\mathbf{n}}{\mathbf{N}}(\mathbf{P}_1 - \mathbf{P}_0)$$

where

P_t Estimate of a given population at a given inter censual year t.
P_0 Population in the previous census.
N Difference between two census years and "**n**" is the number of years between the given year and the previous census year.
P_1 Population in the succeeding census.

Example: The population of the island was 124 million in the census 1991. They conduct decennial census enumeration. In the year 2001, the value rose to 234 million. Calculate the total population of the island in the year 1996.

$$P_t = P_0 + \frac{n}{N}(P_1 - P_0)$$

$$P_{1996} = 124 + \frac{5}{10}(234 - 124) = 179.$$

The estimate for the population value for the year 1996 is 179.

Example: A population estimate was to be calculated to compute the crude birth rate for a popular city, North Charleston in South Carolina for the year 1979.

Birth in country in 1999 =3561.
1970 census count =181,935.
1980 census count =223,814.
Population projection for the year 1980

$$= 181{,}935 + \frac{111}{120}(223{,}814 - 181{,}935)$$

$$= 181{,}935 + 0.925 * 41879 = 220{,}673$$

2. Intercensus estimation with mid-year Population.

This estimate is vital for the socioeconomic planning programs. In this method, we assume that the population growth is linear and the yearly change in the population sizes are equal. Intercensus estimation with mid-year population involves estimating the population of a region or country for a specific time period between two official census counts. This estimation is usually done for the mid-year point between the two census years. It is important for various planning purposes, resource allocation, and policy decisions.

Mid-year Population = Population at the last official census * (1 + Average Annual Growth Rate) ^ Number of years.

where:

Population at the last official census: The population recorded during the last official census.

Average Annual Growth Rate: The average annual growth rate of the population between the two census years. It is calculated as (Population at the next official census/Population at the last official census) ^ (1/Number of years) −1.

Number of years: The time difference (in years) between the mid-year point you want to estimate and the last official census.

- Remember to use consistent units for population and time to get accurate results.

Example 1: Suppose Country X conducted its last official census on January 1, 2020, and recorded a population of 50 million. The next official census is scheduled for January 1, 2025. We want to estimate the population of Country X for July 1, 2023, using the intercensus method.

Step 1: Calculate the time interval between the two census years: The time between January 1, 2020, and January 1, 2025, is 5 years.

Step 2: Determine the time difference between the last census and the mid-year population estimate: The time between January 1, 2020, and July 1, 2023, is approximately 3.5 years.

Step 3: Calculate the average annual growth rate: The average annual growth rate can be calculated using the formula: Average Annual Growth Rate =(Population at 2025 census/Population at 2020 census) ^ (1/Number of years) −1.

Average Annual Growth Rate =(Population at 2025 census/Population at 2020 census) ^ (1/5) −1 Average Annual Growth Rate =(50 million/50 million)^(1/5) − 1 Average Annual Growth Rate =1^0.2 −1 Average Annual Growth Rate =0.1487 or 14.87%

Step 4: Estimate the mid-year population for 2023: To estimate the population for July 1, 2023, we will use the formula: Mid-year Population =Population at 2020 census * (1 +Average Annual Growth Rate) ^ Number of years.

Mid-year Population =50 million * (1 +0.1487) ^ 3.5 Mid-year Population = 50 million * 1.532 Mid-year Population ≈76.6 million.

So, the estimated population of Country X for July 1, 2023, using the intercensus method with mid-year population, is approximately 76.6 million.

Example 2: Suppose Country Y conducted its last official census on April 1, 2019, and recorded a population of 80 million. The next official census is scheduled for April 1, 2024. We want to estimate the population of Country Y for October 1, 2022, using the intercensus method.

Step 1: Calculate the time interval between the two census years: The time between April 1, 2019, and April 1, 2024, is 5 years.

Step 2: Determine the time difference between the last census and the mid-year population estimate: The time between April 1, 2019, and October 1, 2022, is approximately 3.5 years.

Step 3: Calculate the average annual growth rate: The average annual growth rate can be calculated using the formula: Average Annual Growth Rate =(Population at 2024 census/Population at 2019 census) ^ (1/Number of years) −1.

Average Annual Growth Rate =(Population at 2024 census/Population at 2019 census) ^ (1/5) −1 Average Annual Growth Rate =(80 million/80 million) ^ (1/5) −1 Average Annual Growth Rate =1^0.2 −1 Average Annual Growth Rate =0 or 0%

Step 4: Estimate the mid-year population for 2022: To estimate the population for October 1, 2022, we will use the formula: Mid-year Population =Population at 2019 census * (1 +Average Annual Growth Rate) ^ Number of years.

Mid-year Population =80 million * (1 +0) ^ 3.5 Mid-year Population =80 million * 1 Mid-year Population =80 million.

So, the estimated population of Country Y for October 1, 2022, using the intercensus method with mid-year population, is approximately 80 million.

In this example, the average annual growth rate is 0% because the population remained the same between the two census years. This can happen if there is no significant population change or if the population increase and decrease balanced out during the period.

3. Compound growth rate formula

The compound growth rate formula for intercensal data is used to calculate the average annual growth rate of a population between two census years. This formula helps estimate the rate at which the population is growing or declining over a specific period.

The formula for compound growth rate (CGR) is as follows:

CGR =(Population at the end of the period/Population at the beginning of the period) ^ (1/Number of years) −1.

where:

- Population at the end of the period: The population recorded during the later census year.
- Population at the beginning of the period: The population recorded during the earlier census year.
- Number of years: The time interval between the two census years (usually in years).

Let us use this formula in the previous Example 1 to calculate the compound growth rate for Country X between January 1, 2020 (last official census) and January 1, 2025 (next official census):

CGR =(Population at 2025 census/Population at 2020 census) ^ (1/5) −1 CGR =(50 million/50 million) ^ (1/5) −1 CGR =1^0.2 −1 CGR =0.1487 or 14.87%

The calculated compound growth rate is approximately 14.87%. This means that, on average, the population of Country X is growing at a rate of about 14.87% per year between the census years of 2020 and 2025.

Example 1: Suppose City A conducted its last official census on January 1, 2010, and recorded a population of 500,000. The next official census is scheduled for January 1, 2025, and it records a population of 800,000. Calculate the compound growth rate for City A between these two census years.

Step 1: Determine the time interval between the two census years: The time between January 1, 2010, and January 1, 2025, is 15 years.

Step 2: Calculate the compound growth rate (CGR): CGR =(Population at the end of the period/Population at the beginning of the period) ^ (1/Number of years) −1 CGR =(800,000/500,000) ^ (1/15) −1 CGR =1.6 ^ (1/15) −1 CGR ≈0.0332 or 3.32%

The calculated compound growth rate for City A between 2010 and 2025 is approximately 3.32%.

Example 2: Suppose Country B conducted its last official census on July 1, 2018, and recorded a population of 12 million. The next official census is scheduled for July 1, 2023, and it records a population of 15 million. Calculate the compound growth rate for Country B between these two census years.

Step 1: Determine the time interval between the two census years: The time between July 1, 2018, and July 1, 2023, is 5 years.

Step 2: Calculate the compound growth rate (CGR): CGR =(Population at the end of the period/Population at the beginning of the period) ^ (1/Number of years) −1 CGR =(15,000,000/12,000,000) ^ (1/5) −1 CGR =1.25 ^ (1/5) −1 CGR ≈ 0.0471 or 4.71%

The calculated compound growth rate for Country B between 2018 and 2023 is approximately 4.71%.

4. Natural Increase and the net migration method

Natural increase and the net migration method are two approaches used in demography and population studies to estimate intercensal data, particularly the population change between two censuses or population counts taken at different points in time. Both methods are essential for understanding population dynamics and planning various public policies. Let us delve into each method:

1. Natural Increase Method: Natural increase refers to the difference between the number of births and the number of deaths in a population over a specific period. To estimate population change using the natural increase method, follow these steps:

Step 1: Calculate the number of births during the intercensal period. Step 2: Calculate the number of deaths during the intercensal period. Step 3: Determine the difference between the number of births and the number of deaths. This gives you the natural increase for the period. Step 4: Add the natural increase to the population at the beginning of the intercensal period to obtain the estimated population at the end of the period.

Population at the end of the period = Population at the beginning + Natural Increase.

2. Net Migration Method: Net migration refers to the difference between the number of people who have moved into an area (in-migration) and the number of people who have moved out of the area (out-migration) over a specific period. To estimate population change using the net migration method, follow these steps:

Step 1: Calculate the number of people who have migrated into the area (in-migration) during the intercensal period. Step 2: Calculate the number of people who have migrated out of the area (out-migration) during the intercensal period. Step 3: Determine the difference between in-migration and out-migration. This gives you the net migration for the period. Step 4: Add the net migration to the population at the beginning of the intercensal period to obtain the estimated population at the end of the period.

Population at the end of the period = Population at the beginning +Net Migration.

It is important to note that the accuracy of these methods depends on the quality of data collected during censuses and the reliability of demographic statistics, such as birth and death registration systems and migration records.

For more accurate estimates, demographers often combine both the natural increase method and the net migration method to account for all components of population change (births, deaths, and migration). This approach is commonly referred to as the demographic accounting method and provides a comprehensive understanding of how a population changes over time.

Example 1: Natural Increase Method.

Let us consider a hypothetical population in a town between two censuses, where the population at the beginning of the intercensal period (initial population) was 50,000, and the data collected during the period is as follows:

Number of births: 2500 Number of deaths: 1000.

Step 1: Calculate the natural increase. Natural Increase =Number of births − Number of deaths Natural Increase =2500 −1000 Natural Increase =1500.

Step 2: Calculate the estimated population at the end of the period. Population at the end of the period =Population at the beginning +Natural Increase Population at the end of the period =50,000 +1500 Population at the end of the period =51,500.

So, based on the natural increase method, the estimated population at the end of the intercensal period is 51,500.

Example 2: Natural Increase Method.

Let us consider another hypothetical population, this time with different data:

Population at the beginning of the intercensal period: 800,000 Number of births: 20,000 Number of deaths: 8000.

Step 1: Calculate the natural increase. Natural Increase =Number of births − Number of deaths Natural Increase =20,000 −800 Natural Increase =12,000.

Step 2: Calculate the estimated population at the end of the period. Population at the end of the period =Population at the beginning +Natural Increase Population at the end of the period =800,000 +12,000 Population at the end of the period = 812,000.

So, based on the natural increase method, the estimated population at the end of the intercensal period is 812,000.

Example 1: Net Migration Method.

Now, let us work out an example using the net migration method. Consider the following data for a city:

Population at the beginning of the intercensal period: 300,000 Number of people who migrated into the city (in-migration): 25,000 Number of people who migrated out of the city (out-migration): 12,000.

Step 1: Calculate the net migration. Net Migration =In-migration −Out-migration Net Migration =25,000 −12,000 Net Migration =13,000.

Step 2: Calculate the estimated population at the end of the period. Population at the end of the period =Population at the beginning +Net Migration Population at the end of the period =300,000 +13,000 Population at the end of the period = 313,000.

So, based on the net migration method, the estimated population at the end of the intercensal period is 313,000.

Example 2: Net Migration Method.

Let us consider another example with different data:

Population at the beginning of the intercensal period: 1000,000 Number of people who migrated into the city (in-migration): 50,000 Number of people who migrated out of the city (out-migration): 30,000.

Step 1: Calculate the net migration. Net Migration =In-migration −Out-migration Net Migration =50,000 −30,000 Net Migration =20,000.

Step 2: Calculate the estimated population at the end of the period. Population at the end of the period =Population at the beginning +Net Migration Population at the end of the period =1,000,000 +20,000 Population at the end of the period = 1,020,000.

So, based on the net migration method, the estimated population at the end of the intercensal period is 1,020,000.

In real-world scenarios, demographers often combine both methods to get a more accurate estimate of population change by considering both natural increase and net migration.

Reference

Weden, M. M., Peterson, C. E., Miles, J. N., & Shih, R. A. (2015). Evaluating Linearly Interpolated Intercensal Estimates of Demographic and Socioeconomic Characteristics of U.S. Counties and Census Tracts 2001–2009. *Population research and policy review*, *34*(4), 541. https://doi.org/10.1007/s11113-015-9359-8.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4624462/table/T1/?report=objectonly

4.6 Rates and Ratios of Vital Event

Consider a population study between time periods [t_1 and t_2], and there are n1 people born, n2 people married, n3 number of people dead, and n4 number of people divorced. To turn these facts into some information with utility, we transform them into rates and ratios.

Rates refer to "per thousand people," and ratios are values used for comparison. The most frequently used ratio in vital statistics is the sex ratio, employment ratio, and literacy ratio.

1. Birth Rate (Rate): The birth rate refers to the number of births per thousand people in the population between time periods [t1 and t2]. It can be calculated as follows: Birth Rate =(Number of births/Total population) * 1000

Example: If there were 500 births in a population of 10,000 between t1 and t2, the birth rate would be: Birth Rate =(500/10,000) * 1000 =50 births per thousand people.

2. Marriage Rate (Rate): The marriage rate represents the number of marriages per thousand people in the population between time periods [t1 and t2]. It can be

calculated as follows: Marriage Rate =(Number of marriages/Total population) * 1000

Example: If there were 300 marriages in a population of 15,000 between t1 and t2, the marriage rate would be: Marriage Rate =(300/15,000) * 1000 =20 marriages per thousand people.

3. Death Rate (Rate): The death rate is the number of deaths per thousand people in the population between time periods [t1 and t2]. It can be calculated as follows: Death Rate =(Number of deaths/Total population) * 1000

Example: If there were 200 deaths in a population of 8000 between t1 and t2, the death rate would be: Death Rate =(200/8000) * 1000 =25 deaths per thousand people.

4. Divorce Rate (Rate): The divorce rate refers to the number of divorces per thousand married people in the population between time periods [t1 and t2]. It can be calculated as follows: Divorce Rate =(Number of divorces/Number of married people) * 1000

Example: If there were 50 divorces among 400 married people between t1 and t2, the divorce rate would be: Divorce Rate =(50/400) * 1000 =125 divorces per thousand married people.

5. Sex Ratio (Ratio): The sex ratio is a ratio of the number of males to females in the population. It helps in understanding the gender distribution. Sex Ratio = (Number of males/Number of females)

Example: If there are 600 males and 700 females in the population, the sex ratio would be: Sex Ratio =600/700 $\approx$0.86.

6. Employment Ratio (Ratio): The employment ratio represents the proportion of employed individuals in the population. It is usually expressed as a percentage. Employment Ratio =(Number of employed individuals/Total population) * 100

Example: If there are 400 employed individuals in a population of 2000, the employment ratio would be: Employment Ratio =(400/2000) * 100 =20%

7. Literacy Ratio (Ratio): The literacy ratio is the proportion of literate individuals in the population. It is also typically expressed as a percentage. Literacy Ratio = (Number of literate individuals/Total population) * 100

Example: If there are 1200 literate individuals in a population of 1800, the literacy ratio would be: Literacy Ratio =(1200/1800) * 100 =66.67%

These rates and ratios provide valuable information about the population, enabling comparisons and insights into various aspects of the study.

4.7 Mortality and Death Rates

Mortality and death rates are important demographic indicators used to measure the occurrence of deaths within a specific population over a particular period. While the terms are related, they have slightly different meanings:

1. Mortality Rate: Mortality rate refers to the number of deaths in a given population within a specific time frame, typically expressed as the number of deaths per 1000 or 100,000 individuals in the population. Mortality rates can be calculated for various groups, such as age-specific mortality rate (deaths within a specific age group), infant mortality rate (deaths of infants under one year old), or Maternal Mortality Rate (deaths related to pregnancy and childbirth).

Mortality rates are essential for assessing the health status of a population, identifying patterns of diseases, and measuring the effectiveness of healthcare interventions.

2. Death Rate: The death rate, also known as the Crude Death Rate, is a broader measure that represents the total number of deaths within a population per year, usually per 1000 or 100,000 people. It is calculated by dividing the number of deaths in a given year by the mid-year population and multiplying by a constant (e.g., 1000 or 100,000).

The death rate provides a general overview of mortality in a population and is often used to compare mortality levels across different regions or countries.

It is important to note that both mortality and death rates are essential statistics for public health and epidemiology, helping researchers and policymakers understand health trends, identify risk factors, and formulate strategies to improve overall population health. These rates can also vary significantly based on factors such as access to healthcare, lifestyle behaviors, infectious disease prevalence, socioeconomic status, and environmental conditions.

- Are death rates and mortality rate the same?

A mortality rate is a measurement of how frequently people die within a given population during a certain period of time. It just depends on what you want to measure—illness or death—because mathematically, morbidity and mortality metrics are frequently equivalent. The population size at the mid-point of the time period is typically used as the denominator when mortality rates are calculated using vital data (such as counts of death certificates).

Both mortality and death rates are essential for understanding the health status of a population, identifying potential health issues, and evaluating the impact of public health interventions. These rates can vary significantly across different regions and can be used to compare the health outcomes of different populations.

Example 1: Calculating Mortality Rate.

Suppose we want to calculate the mortality rate for a specific age group (20–29 years old) in a town over the course of one year. In that period, there were 15 deaths among individuals aged 20–29 years.

Step 1: Identify the relevant information:

- Number of deaths in the age group (20–29): 15

Step 2: Calculate the mortality rate: Mortality Rate =(Number of deaths/Total population in the age group) $\times$1000.

If the total population of the age group (20–29) is 5000, then the mortality rate would be: Mortality Rate =(15/5000) $\times$1000 =3 deaths per 1000 individuals in the age group.

Example 2: Calculating Death Rate.

Now, let us calculate the death rate for an entire country over the course of one year. During that year, the country's total population was 10 million, and the number of deaths recorded was 100,000.

Step 1: Identify the relevant information:

- Total number of deaths in the country: 100,000
- Total population of the country: 10,000,000

Step 2: Calculate the death rate: Death Rate = (Number of deaths/Total population) $\times$1000.

Death Rate =(100,000/10,000,000) $\times$1000 =10 deaths per 1000 individuals in the country.

In this example, the death rate for the country is 10 deaths per 1000 individuals.

4.7.1 Crude Mortality Rate or the Crude Death Rate

The Crude Mortality Rate (CMR) or Crude Death Rate (CDR) is a vital demographic indicator used to measure the number of deaths in a population per unit of time, usually expressed as the number of deaths per 1000 or 100,000 people in the population. It provides a general overview of the mortality level in a particular region or country and is commonly used in public health, epidemiology, and demography.

The formula to calculate the Crude Mortality Rate is:

Crude Mortality Rate (CMR) =(Number of Deaths/Total Population) ×Multiplier.

where:

- Number of Deaths: The total number of deaths in the specified population during a given period (e.g., a year).
- Total Population: The estimated or actual population of the area or country at the mid-point of the specified period.
- Multiplier: To convert the CMR to a per 1000 or 100,000 scale, a multiplier is used. If you want to express it per 1000, the multiplier is 1000; if you want to express it per 100,000, the multiplier is 100,000.

Here are a couple of solved examples to illustrate how to calculate the Crude Mortality Rate:

Example 1: Let us say we have a small town with a population of 10,000 people, and during the year 2022, a total of 50 deaths occurred.

CMR =(50/10,000) ×1000 CMR =0.005 ×1000 CMR =5 deaths per 1000 people.

Example 2: Now, consider a larger city with a population of 500,000. During the same year 2022, there were 1800 deaths.

CMR =(1800/500,000) ×1000 CMR =0.0036 ×1000 CMR =3.6 deaths per 1000 people.

It is important to note that the Crude Mortality Rate does not take into account the age distribution of the population, which can significantly impact mortality rates. To get more detailed insights into mortality patterns, age-specific mortality rates and other measures are used. Nonetheless, the Crude Mortality Rate remains a valuable and straightforward tool for comparing mortality levels between different regions or for tracking changes in mortality over time.

4.7.2 Cause-Specific Mortality Rate and Age-Specific Mortality Rate

The mortality rate for a population attributable to a certain cause is known as the Cause-Specific Mortality Rate. The number of fatalities linked to a certain cause is

the numerator. The population's size at the period's halfway point continues to serve as the denominator. The percentage is often given per 100,000 people.

A mortality rate that is only applicable to a certain age group is called an Age-Specific Mortality Rate. The denominator is the number of people in that age group in the population, while the numerator is the number of deaths in that age group.

4.7.3 Neonatal Mortality Rate

Newborn mortality rate is also known as neonatal mortality rate. Neonatal mortality specifically refers to the number of deaths occurring in the first 28 days (0–27 days) of life per 1000 live births in a given year. This indicator focuses on the early stage of infancy when newborns are most vulnerable and at higher risk of mortality. Neonatal mortality is an essential measure to assess the health and well-being of newborns and to track progress in improving maternal and child health outcomes.

Typically, the newborn mortality rate is determined annually. Because it accounts for both the mother's and the baby's health throughout pregnancy and the first year after, it is a widely used indicator of health status. Access to prenatal care, the prevalence of prenatal maternal health behaviors (like alcohol or tobacco use and proper nutrition during pregnancy, etc.), postnatal care and behaviors (like childhood immunizations and proper nutrition), sanitation, and infection control are just a few examples of the many factors that have an impact on the health of the mother and child.

The neonatal period includes the first 28 days after birth but not beyond. Therefore, the number of infant deaths under the age of 28 days constitutes the numerator of the neonatal mortality rate. The number of live births reported within the same time period serves as the denominator for both the neonatal mortality rate and the infant mortality rate. The standard way to express the neonatal death rate is per 1000 live births.

As per the last knowledge update in September 2021, the global newborn mortality rate was declining but still remained a significant public health concern. The newborn mortality rate refers to the number of deaths of infants aged 0 to 28 days per 1000 live births in a given year. It is an important indicator of the overall health and healthcare system of a country.

According to data from the World Health Organization (WHO) and the United Nations Children's Fund (UNICEF), the global newborn mortality rate was estimated to be around 18 deaths per 1000 live births in 2019. This means that approximately 2.8 million newborns died in the first month of life worldwide that year.

It is crucial to note that newborn mortality rates can vary significantly between different countries and regions. In high-income countries, the rates tend to be much lower compared to low-income and resource-limited countries, where access to quality healthcare and other factors can influence the survival of newborns.

It is recommended to check more recent sources, such as the WHO or UNICEF websites, for the latest data on newborn mortality rates.

The period from 28 days of age up to but excluding 1 year of age is known as the post-neonatal phase. Therefore, the number of fatalities among children between the ages of 28 days and but not including 1 year over a specific time period constitutes the numerator of the post-neonatal mortality rate. The number of live births reported within the same time period serves as the denominator. The standard way to express the post-neonatal death rate is per 1000 live births.

It is often used in discussions related to infant mortality to distinguish between neonatal mortality (deaths that occur within the first 28 days of life) and post-neonatal mortality (deaths that occur between 28 days and one year of age).

To summarize:

- Neonatal mortality: Deaths that occur within the first 28 days (0–27 days) of life.
- Post-neonatal mortality: Deaths that occur between 28 days and one year (364 days) of age.

Together, neonatal and post-neonatal mortality rates provide a comprehensive view of infant mortality, which is an essential indicator for assessing the health and survival of infants during their first year of life.

4.7.4 Maternal Mortality Rate (M.M.R)

Maternal Mortality Rate (MMR) refers to the number of maternal deaths that occur during pregnancy or within 42 days of termination of pregnancy (regardless of the duration and site of the pregnancy), per 100,000 live births in a given time period. It is a crucial indicator of the overall health and well-being of women in a particular country or region and reflects the effectiveness of a healthcare system in providing proper maternal care and reducing preventable deaths related to pregnancy and childbirth.

High Maternal Mortality Rates are often associated with inadequate access to quality healthcare, limited availability of skilled birth attendants, poor nutrition, and socioeconomic factors. Conversely, lower Maternal Mortality Rates are generally observed in regions with better healthcare infrastructure, prenatal care, access to skilled healthcare professionals, and a focus on maternal health.

It is important to track and reduce Maternal Mortality Rates to improve maternal health and promote safer pregnancies and childbirth experiences for women globally. The World Health Organization (WHO) and various other organizations and governments regularly collect and publish data on Maternal Mortality Rates to assess progress and target interventions to address maternal health issues.

MMR =(Number of Maternal Deaths/Number of Live Births) $\times$100,000.

Let us say we have the following data for a specific region in a given year:

Number of maternal deaths =50 Number of live births =10,000.

Now, we can calculate the MMR using the formula:

M.M.R =(50/10,000) $\times$100,000 =500.

Therefore, the Maternal Mortality Rate (MMR) for this specific region in the given year is 500 maternal deaths per 100,000 live births. This means that, on average, there were 500 maternal deaths for every 100,000 live births during that year in that particular region.

4.7.5 Sex-Specific Mortality Rate

The sex-specific mortality rate is a measure that calculates the number of deaths in a specific sex group (e.g., males or females) per a given population's size. It is typically expressed as the number of deaths per 1000 individuals of that sex in a given time period. To calculate the sex-specific mortality rate, you will need the number of deaths and the population size for the specific sex group.

Example: Suppose we have the following information for a population in a specific time period:

- Number of deaths among males: 400.
- Total population of males: 15,000.

To calculate the sex-specific mortality rate for males:

Step 1: Divide the number of deaths by the total population of males. Mortality rate =Number of deaths/Total population of males Mortality rate =400/15,000.

Step 2: Calculate the rate per 1000 individuals by multiplying the result from Step 1 by 1000. Mortality rate per 1000 males =(400/15,000) * 1000 Mortality rate per 1000 males =26.67.

Step 3: Round the mortality rate to an appropriate number of decimal places, if needed. Rounded mortality rate per 1000 males =26.7

So, the sex-specific mortality rate for males in this example is approximately 26.7 deaths per 1000 males in the given time period. Remember that this is just an example, and actual mortality rates will vary depending on the population and time frame being studied.

4.7.6 Race-Specific Mortality Rate

Race-specific mortality rate refers to the death rate within a particular racial or ethnic group due to specific causes, such as diseases, accidents, or other health-related factors. This rate is calculated as the number of deaths within a specific racial or ethnic group divided by the total population of that group, multiplied by a constant (usually 1000 or 100,000) to express the rate per 1000 or 100,000 individuals.

Health disparities, including differences in mortality rates, have been observed among different racial and ethnic groups. These disparities can be influenced by various factors, including access to healthcare, socioeconomic status, environmental

factors, lifestyle choices, and historical injustices. Understanding and addressing these disparities are critical to promoting health equity and improving overall public health.

It is essential to note that discussing race-specific mortality rates should be approached with sensitivity, as the concept of race is complex and influenced by social, historical, and political factors. Healthcare professionals and policymakers must be mindful of the potential implications and avoid perpetuating stereotypes or promoting discrimination based on race or ethnicity. Instead, efforts should focus on identifying and addressing the underlying causes of health disparities and ensuring equitable access to healthcare for all individuals, regardless of their racial or ethnic background.

Let us consider a hypothetical situation with two racial groups: Group A and Group B. We will calculate the race-specific mortality rate for each group based on the number of deaths from a specific cause within a given time period.

Example: Time Period: January 1, 2023, to December 31, 2023.

Cause of Death: Heart Disease.

Population of Group A: 50,000 Number of Deaths from Heart Disease in Group A: 500.

Population of Group B: 70,000 Number of Deaths from Heart Disease in Group B: 400.

Step 1: Calculate the mortality rate per 100,000 individuals for each group.

Mortality Rate for Group A: (500/50,000) * 100,000 =1000 deaths per 100,000 individuals.

Mortality Rate for Group B: (400/70,000) * 100,000 = 571.43 deaths per 100,000 individuals.

Step 2: Compare the race-specific mortality rates.

In this example, the race-specific mortality rate for Group A due to heart disease is 1000 deaths per 100,000 individuals, while the rate for Group B is 571.43 deaths per 100,000 individuals.

Note: that this is a simplified hypothetical example for illustrative purposes. In reality, race-specific mortality rates are calculated using more comprehensive data from various sources and over extended periods to identify trends and patterns accurately. Additionally, considering more causes of death and a broader range of racial or ethnic groups would provide a more comprehensive analysis of health disparities. Health authorities and researchers use such data to develop targeted interventions and policies aimed at reducing health inequalities and promoting health equity among different populations.

4.7.7 Age-Specific Death Rates

$$\text{A.S.D.R} = \frac{\text{Number of deaths among people in a particular age}}{\text{Total number of people in the particular age}} * 1000$$

Example: Consider the number of deaths and population in the 65–69 age group during a particular year in a hypothetical country.

Suppose we have the following data:

- Number of deaths in the 65–69 age group: 500.
- Population of the 65–69 age group: 25,000.

We will calculate the age-specific death rate (ASDR) for this age group using the formula mentioned earlier:

$$\text{ASDR} = (\text{Number of deaths in the age group}/\text{Population of the age group}) \times \text{K}$$

Let us assume we want to express the death rate per 1000 population (K =1000).

$$\text{ASDR} = (500/25{,}000) \times 1000$$
$$\text{ASDR} = 20$$

The age-specific death rate for the 65–69 age group is 20 deaths per 1000 population. This means that, during the specified time period, there were 20 deaths for every 1000 people within the 65–69 age group.

Note: Age-specific death rates can vary significantly between different age groups and populations. By calculating and analyzing these rates for various age groups, policymakers and researchers can gain insights into the health status and mortality patterns within different segments of the population, which can inform targeted public health interventions and policy decisions.

Example 2: Consider the following data (Table 4.2).

Did age-specific death rates in 2020 change from 2019 for those aged 1 year and over?

From 2019 to 2020, death rates increased for each age group 15 years and over. Rates increased 20.8% for age group 15–24 (from 69.7 deaths per 100,000 population in 2019 to 84.2 in 2020), 23.8% for 25–34 (128.8 to 159.5), 24.5% for 35–44 (199.2 to 248.0), 20.7% for 45–54 (392.4–473.5), 17.6% for 55–64 (883.3–1038.9), 17.4% for 65–74 (1764.6 to 2072.3), 16.0% for 75–84 (4308.3–4997.0), and 15.0% for 85 and over (13,228.6–15210.9) (Fig. 4.3). Rates for age groups 1–4 and 5–14 did not change significantly from 2019 to 2020 (Fig. 4.4).

Statistically significant increase in age-specific death rate from 2019 to 2020 NOTES: Rates are plotted on a logarithmic scale. Source: National Centre for Health Statistics, National Vital Statistics System, Mortality.

Table 4.2 Number of deaths and death rates for ages 1 year and over, United States 2019 and 2020

Age group (years)	2019		2020	
	Number	Ratel	Number	Rate[a]
1–4	3676	23.3	3529	22.7
5–14	5497	13.4	5623	13.7
15–24	29,771	69.7	35,816	84.2
25–34	59,178	128.8	73,486	159.5
35–44	82,986	199.2	104,490	248.0
45–54	160,393	392.4	191,142	473.5
55–64	374,937	883.3	440.549	1038.9
65–74	555,559	1764.6	674,507	2072.3
75–84	688,027	4308.3	822,084	4997.0
85 and over	873,746	13,228.6	1,012,805	15,210.9

Source: National center for health statistics, national vital statistics system, mortality
[a] Deaths per 100,000 population

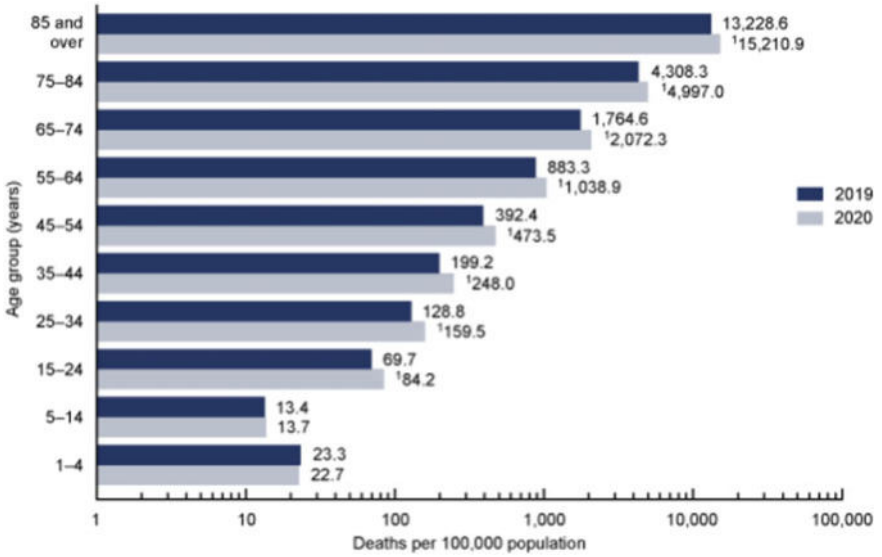


Fig. 4.4 Bar graph for deaths 100,000 population in year 2019 and 2020 categorized based on age

4.7.8 Standardized Death Rates

Standardized death rates, also known as age-standardized death rates or age-adjusted death rates, are a statistical measure used to compare mortality rates between different populations or across different time periods, while accounting for differences in age

distributions. Since the risk of death generally increases with age, comparing raw death rates between populations with different age structures can be misleading.

To calculate standardized death rates, the age-specific death rates in the population of interest are adjusted to a standard population with a fixed age distribution. This standardization process allows for a more meaningful and fair comparison of death rates, as if the populations being compared had the same age distribution.

The steps to calculate age-standardized death rates are as follows:

1. Obtain age-specific death rates: Gather data on the number of deaths for different age groups within the population of interest. For example, you might have death rates for age groups 0–4, 5–9, 10–14, and so on.
2. Obtain the standard population: Choose a standard population with a fixed age distribution, such as the World Health Organization (WHO) standard population or another suitable reference population. The WHO standard population is designed to be broadly representative of the world population.
3. Calculate the weighted average: Calculate the weighted average of the age-specific death rates, using the standard population as the weights. The formula is usually given as:

$$\text{Standardized Death Rate} = \Sigma \left(\begin{array}{l} \text{Population} - \text{specific Death Rate\_i} \\ * \text{Standard Population Proportion\_i} \end{array} \right)$$

where i refers to the age group and the summation is performed over all age groups.

4. Standardize the rate: The final step is to express the weighted average as a rate per a specified unit of the standard population (e.g., per 1000 or 100,000 people) to obtain the standardized death rate.

Standardized death rates are particularly useful when comparing mortality across regions or countries with different age structures, as they allow for more accurate comparisons and facilitate better understanding of health disparities and trends.

It is important to note that the process of standardization is just one aspect of understanding mortality patterns. When interpreting death rates, it is essential to consider other factors such as the cause of death, underlying health conditions, access to healthcare, and socioeconomic factors that may influence mortality in different populations (Fig. 4.5).

Example 1: Consider the following data (Table 4.3):

Data Source: Web-based Injury Statistics Query and Reporting System (WISQARS) [online database] Atlanta; National Center for Injury Prevention and Control. Available from: https://www.cdc.gov/injury/wisqars.

The table provided the number of deaths from all causes and from accidents (unintentional injuries) by age group in the United States in 2002. Review the following rates. Determine what to call each one, then calculate it using the data provided.

a. Unintentional-injury-specific mortality rate for the entire population

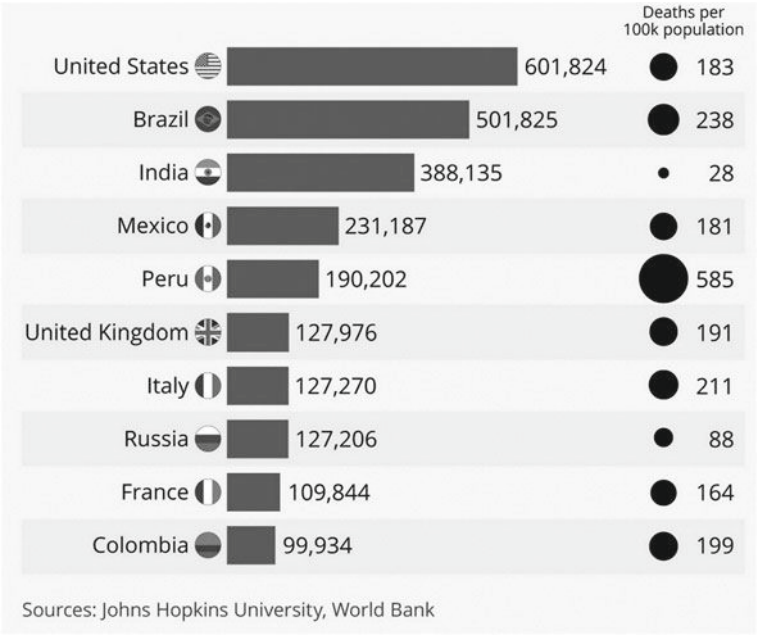


Fig. 4.5 Bar graph for number of deaths per 100k population categorized based on countries. https://www.statista.com/chart/24258/countries-with-the-highest-number-of-covid-19-deaths/

Table 4.3 All cause and unintentional injury mortality and estimated population by age group, for males alone—United States, 2002

Age group (years)	All races, both_sexes			All races, Males		
	All_causes	Unintentional injuries	Estimated population (1000)	all causes	Unintentional injuries	Estimated population (1000)
0–4	32,892	2587	19,597	18,523	1577	10,020
5–14	7150	2718	41,037	4198	1713	21,013
15–24	33,046	15,412	40,590	24,416	11,438	20,821
25–34	41,355	12,569	39,928	28,736	9635	20,203
35–44	91,140	16,710	44,917	57,593	12,012	22,367
45–54	172,385	14,675	40,084	107,722	10,492	19,676
55–64	253,342	8345	26,602	151,363	5781	12,784
65 + years	1,811,720	33,641	35,602	806,431	16,535	14,772
Not stated	357	85	0	282	74	0
Total	2,443,387	106,742	288,357	1,199,264	69,257	141,656

This is a cause-specific mortality rate given as:

$$= \frac{\text{Number of unintentional injury deaths in the entire population}}{\text{Estimated midyear population}} * 100{,}000$$
$$= (106{,}742/288{,}357{,}000) \times 100{,}000$$
$$= 37.0 \text{ unintentional-injury-related deaths per } 100{,}000 \text{ population}$$

b. All-cause mortality rate for 25–34 year olds

This is an age-specific mortality rate.

$$\text{Rate} = \frac{\text{Number of deaths from all the causes among people aged between } 25-34 \text{ years}}{\text{Estimated midyear population of } 25-34 \text{ year}} * 100{,}000$$
$$= 103.6 \text{ deaths per } 100{,}00025-34 \text{ year olds}$$

c. All-cause mortality among males

This is a sex-specific mortality rate.

$$\text{Rate} = \frac{\text{Number of deaths from all causes among males}}{\text{Estimated midyear population of males}} * 100{,}000$$
$$= (1{,}199{,}264/141{,}656{,}000\ ) \times 100{,}000$$
$$= 846.6 \text{ deaths per } 100{,}000 \text{ males}$$

d. Unintentional-injury-specific mortality among 25–34 year old males

This is a cause-specific, age-specific, and sex-specific mortality rate.

$$\text{Rate} = \frac{\text{Number of unintentional injury deaths among } 25-34 \text{ year old males}}{\text{Estimated midyear population of } 25-34 \text{ year old males}} * 100{,}000$$
$$= (9{,}635/20{,}203{,}000\ ) \times 100{,}000.$$
$$= 47.7 \text{ unintentional-injury-related deaths per } 100{,}00025-34 \text{ year olds}$$

Example 2: Compare Town A and Town B based on their standardized death rates, which town is healthier? When standardized population values are given (Table 4.4).

Solution: For Town A (Table 4.5).

$$\text{S.D.R(Town A)} = \frac{\sum \text{PA}}{\sum \text{P}} = \frac{190{,}915.92}{17900} = \mathbf{10.665}$$

For Town B (Table 4.6).

Table 4.4 Population of Town A and Town B with standardized population values given

Age group (years)	Town A		Town B		Standard population
	Population	Death	Population	Death	
0–10 years	5603	42	3781	28	2500
10–25 years	12,674	58	20,987	105	3700
25–60 years	6764	61	4532	54	6800
60 years and above	8765	168	3378	67	4900

Table 4.5 Calculation of standardized death rate for Town A

Age group (years)	Town A				
	Population	Death	Standard population	A.S.D.R	P.A
0–10 years	5603	42	2500	7.50	18,739.96
10–25 years	12,674	58	3700	4.58	16,932.30
25–60 years	6764	61	6800	9.02	61,324.66
60 years and above	8765	168	4900	19.17	93,919.00
Total				17,900	190,915.92

Table 4.6 Calculation of standardized death rate for Town B

Age group (years)	Town B		A.S.D.R	Standardised population (P)	PA
	Population	Death			
0–10 years	3781	28	7.41	2500	18,513.62
10–25 years	20,987	105	5.00	3700	18,511.46
25–60 years	4532	54	11.92	6800	81,023.83
60 years and above	3378	67	19.83	4900	97,187.69
Total				17,900.00	215,236.60

$$\text{S.D.R (Town A)} = \frac{\sum \text{PA}}{\sum \text{P}} = \frac{215{,}236.60}{17{,}900} = \mathbf{12.023}$$

Inference: Town A is healthier than Town B since, the S.T.D.R(A) < S.T.D.R(B).

Example: Keeping Town A as a base for comparison, compare both towns and conclude on their health based on their S.T.D.R value (Table 4.7).

Solution

Let us consider Town A (Table 4.8).

Table 4.7 Deaths from Town A and Town B

Age group (years)	Town A		Town B	
	Population	Death	Population	Death
0–10 years	15,903	34	18,764	40
10–25 years	12,454	65	19,764	134
25–60 years	7784	87	8799	65
60 years and above	8765	190	9087	99

Table 4.8 Calculation of S.D.R for town A

Age group (years)	Town A			
	Population	Death	A.S.D.R	Population *A.S.D.R
0–10 years	15,903	34	2.14	34,000
10–25 years	12,454	65	5.22	65,000
25–60 years	7784	87	11.18	87,000
60 years and above	8765	190	21.68	1,90,000
Total	44,906	376	40.21106587	3,76,000

$$\text{S.D.R(Town A)} = \frac{\sum \text{PA}}{\sum \text{P}} = \frac{376{,}000}{44{,}906} = \mathbf{8.37}$$

For Town B (Table 4.9),

$$\text{S.D.R (Town B)} = \frac{\sum \text{PB}}{\sum \text{P}} = \frac{478{,}000}{56414} = \mathbf{8.47}$$

Inference: Town A is healthier than Town B since, the S.T.D.R(A) < S.T.D.R(B).

Example: In 2001, a total of 15,556 homicide deaths occurred among males and 4853 homicide deaths occurred among females. The estimated 2001 mid-year populations for males and females were 139,913,000 and 144,984,008, respectively.

Table 4.9 Calculation of S.D.R for Town B

Age group (years)	Town B			
	Population	Death	ASDR	Population *A.S.D.R
0–10 years	18,764	40	2.13	40,000
10–25 years	19,764	134	6.78	1,34,000
25–60 years	8799	65	7.39	65,000
60 years and above	9087	99	16.30	2,39,000
Total	56,414	338	32.60	4,78,000

a. Calculate the homicide-related death rates for males and for females.
b. What type(s) of mortality rates did you calculate in Question 1?
c. Calculate the ratio of homicide-mortality rates for males compared to females.
d. Interpret the rate you calculated in Question 3 as if you were presenting information to a policymaker.

1. Homicide-related death rate for males.

$$= (\text{Number of homicide deaths among males}/\text{Total male population}) \times 100{,}000$$
$$= 15{,}556/139{,}913{,}000 \times 100{,}000 = \mathbf{11.1183}$$

This means that 11.1183 homicide deaths per 100,000 population among males.
Homicide-related death rate for females.

$$= (\text{Number of homicide deaths among females}/\text{Total female population}) \times 100{,}000$$
$$= 4853/144{,}984{,}008 \times 100{,}000 = \mathbf{3.347}$$

This means that 3.347 homicide deaths / 100,000 population among females.

2. These are **cause-specific** and **sex-specific** mortality rates.
3. Homicide-mortality rate ratio is given as

$$= \text{Homicide death rate of males}/\text{Homicide death rate of females}$$
$$= 11.1183/3.347 = \mathbf{3.321}$$

4. Because the homicide rate among males is higher than the homicide rate among females, specific intervention programs are required to be set up to target males and females differently.

Note: Mortality rates can be of various combinations of categories based on gender, skills, education levels, types of diseases, stages of pain, age, etc. Most of the researchers keep age as a base of their studies. Few interesting facts on mortality rates are as follows: For example,

- 9 out of every 10 maternal deaths in Asia–Pacific occur in just 12 countries. The facts stated in an article from UNFPA, United Nations Population Fund, were very astonishing that in the year 2015, 92% of the maternal deaths which is approximately 78,000 occurred in just 12 countries. This was an example of region-based classification of MMR. Countries such as Myanmar, New Guinea, Philippines, and Timor-Leste have over 100 deaths per 100,000 live births. (https://asiapacific.unfpa.org/en/news/maternal-mortality-asia-pacific-5-key-facts)
- OurWorldData.org states that many countries in South Africa have death rates due to HIV/AIDS greater than 100 per 100,000 population. In Mozambique, it was over 200 per 100,000 population, whereas in Europe the rates were less than 1 per 100,000 population. https://ourworldindata.org/hiv-aids#death-rates-are-high-across-sub-saharan-africa

Problem: Consider the following data (Table 4.10):

Problem: The number of new cases and deaths from diphtheria declined dramatically from the 1940s through the 1980s, but remained roughly level at very low levels in the 1990s. The death-to-case ratio was actually higher in the 1980s and 1990s than in 1940s and 1950s. From these data one might conclude that the decline in deaths is a result of the decline in cases, that is, from prevention, rather than from any improvement in the treatment of cases that do occur (Table 4.11).

Proportionate mortality for diseases of the heart, 25–44 years.

$$= (\text{Number of deaths from diseases of heart}/\text{Number of deaths from all causes}) \times 100$$
$$= 16{,}283/128{,}294 \times 100$$
$$= 12.6\%$$

Proportionate mortality for assault (homicide), 25–44 years.

Table 4.10 Deaths due to diphtheria from 1940 to 1999

Decade	Number of new cases	Number of deaths	Death-to-case ratio (*100)
1940–1949	143,497	11,228	7.82 (Given)
1950–1959	23,750	1710	7.20
1960–1969	3679	390	10.60
1970–1979	1956	90	4.60
1980–1989	27	3	11.11
1990–1999	22	5	22.72

Table 4.11 Deaths categorized age-wise

Age group (years)	Deaths	Age mid-point	Years to 65	Y.P.L. L
Total	14,095		291,020	
0–4	12	2.5	62.5	750
5–10	25	10	55	1375
15–24	178	20	45	8010
25–34	1839	30	35	64,365
35–44	5707	40	25	142,675
45–54	4474	50	15	67,110
55–64	1347	60	5	6735
65+	509	–	–	–
Not stated	4	–	–	–
Total	14,095			

= (Number deaths from assault (homicide)/#deaths from all causes) ×100
=7367/128,924
=5.7%

In brief (Table 4.12):

4.8 Birth Rates

Birth rates refer to the number of live births per 1000 individuals in a population over a specified period, typically calculated annually. It is a crucial demographic indicator used to understand population dynamics and can provide insights into a country's social, economic, and health conditions.

High birth rates often occur in less developed countries or regions with limited access to family planning and healthcare services. Factors contributing to high birth rates include a lack of education, limited access to contraceptives, cultural or religious beliefs, and the need for labor in agricultural or traditional economies.

Conversely, low birth rates are typically observed in more developed countries with higher levels of education, better access to healthcare and family planning, urbanization, and changing societal attitudes toward family size and roles. In some cases, low birth rates can lead to concerns about population aging and a declining workforce, which may impact economic productivity and social welfare systems.

Governments and policymakers monitor birth rates closely to plan for future healthcare, education, and infrastructure needs, as well as to assess the potential impacts on the labor force and overall economic growth. Birth rates can also influence discussions about immigration policies, as some countries may use immigration to offset population decline or sustain workforce growth.

It is important to note that birth rates can vary significantly between countries and regions and can change over time due to various social, economic, and political factors. Additionally, the global trend of birth rates has seen a general decline over the past few decades as many countries experience demographic transitions, moving from high birth and death rates to lower and more stable levels.

Birth rates, a key demographic indicator plays a crucial role in shaping the population dynamics of a country or region. Understanding birth rates requires a comprehensive examination of various factors, including social, economic, cultural, and healthcare-related elements. In this article, we delve into the significance of birth rates, their historical trends, and the underlying drivers that influence these rates in different parts of the world.

Historical Overview

Birth rates have been subject to significant fluctuations throughout history. In pre-modern societies, high birth rates were common due to various factors such as agrarian economies, lack of effective contraception methods, and high child mortality

Table 4.12 Brief note about all death rates

Measure	Numerator	Denominator	Per	Advantages	Disadvantages
Crude death rate	Annual deaths	Annual mean population	1000	Considers the entire population, simple to compute	It assumes that the risk of exposure to death is uniform in the population. Hence not suitable for comparison
Cause-specific death rate	Number of deaths assigned to a specific cause during a given interval of time	Number of people exposed to the cause under study or mid-interval population	1000	Ideal for research	Usability is limited
Death-to-cause ratio	Number of deaths assigned to a specific cause during a given interval of time	Number of new cases of same disease reported during the given interval of time	1000	Special use case under cause-specific death rate	Usability is limited
Neonatal mortality rate	Number of deaths among children less than 28 days of age during a given interval of time	Number of live births during the given interval of time	1000	The rates are for a specific predefined purpose. The information required for the computation can be collected easily compared to other ratios	The main drawback is under registration of live births, and often fetal deaths are recorded as infant deaths that leads to incorrect values
Post-neonatal mortality rate	Number of deaths among children 28–364 days of age during a given interval of time	Number of live births during the given interval of time	1000		
Infant mortality rate	Number of deaths among children < 1 year of age, excluding fetal deaths during a given interval of time	Number of live births during the given interval of time	1000		
Maternal mortality rate	Number of women deaths caused due to pregnancy during a given interval of time	Number of live births during the given interval of time	1000		Most of the health information recording systems are inaccurate

rates. With the onset of industrialization and modernization, birth rates began to decline in many parts of the world.

During the mid-twentieth century, the global population witnessed an unprecedented surge, known as the "baby boom," mainly in developed countries after World War II. Advances in healthcare, improved sanitation, and economic prosperity were primary factors contributing to this boom. However, in the latter half of the twentieth century, birth rates started declining steadily in many developed nations, while some developing regions experienced slower declines.

Factors Influencing Birth Rates

1. Economic Development: Economic factors significantly impact birth rates. As societies transition from agrarian to industrial economies, there is typically a decline in birth rates. Industrialization leads to urbanization, increased education, and more opportunities for women outside traditional roles, leading to delayed marriages and lower fertility rates.
2. Education and Women's Empowerment**:** Education, particularly for women, has been linked to reduced birth rates. Educated women tend to marry later and have fewer children as they pursue careers and become financially independent. Education also raises awareness about family planning and the use of contraceptives.
3. Family Planning and Contraception: Access to family planning services and contraceptive methods plays a crucial role in influencing birth rates. Countries with comprehensive family planning programs typically have lower birth rates due to the wider availability of contraceptive options.
4. Cultural and Religious Norms: Cultural and religious beliefs can influence attitudes toward family size and contraceptive use. In some societies, large families may be encouraged, while in others, there may be strong religious or cultural restrictions on contraception.
5. Healthcare and Infant Mortality: Countries with better healthcare systems and lower infant mortality rates tend to have lower birth rates. When child survival is uncertain, families may choose to have more children to ensure that some survive to adulthood.
6. Government Policies: Government policies, such as family planning initiatives, parental leave, and child subsidies, can influence birth rates. Some countries offer incentives to encourage families to have more children, while others promote family planning and small family sizes.

Global Variation in Birth Rates

Birth rates vary significantly across regions and countries. As of the early twenty-first century, several general trends have emerged:

1. Developed Countries: Most developed countries experience relatively low birth rates, often below the replacement level (around 2.1 children per woman). These low rates can lead to population aging and potential labor force shortages.
2. Developing Countries: Many developing countries still have relatively high birth rates, but they have been declining. However, the pace of decline varies widely

between nations, with some experiencing rapid drops, while others see slower changes.
3. Sub-Saharan Africa: This region tends to have the highest birth rates globally. Factors like high child mortality rates, low access to education, and limited healthcare contribute to this trend.
4. Eastern Europe: Some countries in Eastern Europe have experienced sharp declines in birth rates due to economic uncertainties, changing social norms, and emigration.

Implications of Birth Rates

1. Population Growth**:** Birth rates are a significant driver of population growth. Countries with high birth rates experience rapid population growth, while those with low birth rates may see population stagnation or decline.
2. Aging Population: Low birth rates combined with increasing life expectancy contribute to an aging population in many developed countries. This demographic shift poses challenges for social welfare systems and the labor market.
3. Dependency Ratio: Birth rates influence the dependency ratio, which measures the number of non-working individuals (children and the elderly) supported by the working-age population. High birth rates can lead to a larger dependent population, potentially straining resources.
4. Economic Impact: Birth rates can affect economic productivity and consumption patterns. Rapid population growth may lead to increased demand for goods and services, while an aging population might reduce the size of the labor force.

4.8.1 Some Interesting Statistics About Birth Rates

1. Global Birth Rate (2021): The global birth rate was estimated to be approximately 18.5 births per 1000 people in 2021. This rate represents a decline from previous decades but still contributes to global population growth.
2. Niger: Highest Birth Rate**:** Niger, a country in Sub-Saharan Africa, has one of the highest birth rates globally. As of 2021, the birth rate in Niger was around 44 births per 1000 people, reflecting the region's high fertility levels.
3. Singapore: Lowest Birth Rate: Singapore has one of the lowest birth rates in the world. In 2021, the birth rate was approximately 7.9 births per 1000 people. This low rate has implications for the country's aging population and workforce.
4. European Union (EU) Average Birth Rate: The average birth rate in the European Union was around 9.3 births per 1000 people in 2021. Many European countries have experienced declining birth rates and population aging.
5. United States: Below Replacement Level: The United States has experienced a decline in birth rates over the past few decades. As of 2021, the US birth rate was below the replacement level, with approximately 11.8 births per 1000 people.

6. Japan: Aging Population: Japan has been facing one of the most significant challenges with an aging population and declining birth rates. In 2021, the birth rate in Japan was approximately 7.9 births per 1000 people.
7. India: Declining Birth Rates: Despite being one of the most populous countries globally, India has experienced a decline in birth rates in recent years. As of 2021, the birth rate was around 17.9 births per 1000 people.
8. Teenage Birth Rates**:** Teenage birth rates vary widely across countries. For instance, the United States has seen a decline in teenage birth rates in recent years, with approximately 16.7 births per 1000 females aged 15–19 in 2021.
9. Educational Attainment and Birth Rates: There is a clear inverse relationship between women's educational attainment and birth rates. Women with higher levels of education tend to have fewer children. For example, in many European countries, birth rates are lower among women with higher education degrees.
10. Fertility Rate vs. Birth Rate**:** The fertility rate, which measures the average number of children a woman is expected to have during her lifetime, can differ from the birth rate. Countries with below-replacement-level birth rates may still have a total fertility rate above 2.1 due to differences in age structure and mortality rates.

These statistics highlight the diversity in birth rates across countries and regions, with some facing challenges related to high birth rates and rapid population growth, while others are dealing with low birth rates and potential population decline. Understanding these variations is essential for formulating effective population policies and social welfare strategies to address the unique demographic challenges each country faces.

4.8.2 Fertility Rates

Fertility rates, a critical demographic indicator, play a central role in shaping a country's population structure and growth. This measurement represents the average number of children a woman is expected to have during her reproductive years, typically between the ages of 15 and 49. Understanding fertility rates is crucial for governments, policymakers, and researchers as it provides valuable insights into population dynamics, family planning, and social and economic development (Fig. 4.6).

Global Trends in Fertility Rates

Fertility rates have undergone significant changes over the past century. In the early twentieth century, most countries had high fertility rates, with women having an average of five or more children. However, with societal and economic transformations, fertility rates have been declining in many regions since the mid-twentieth century.

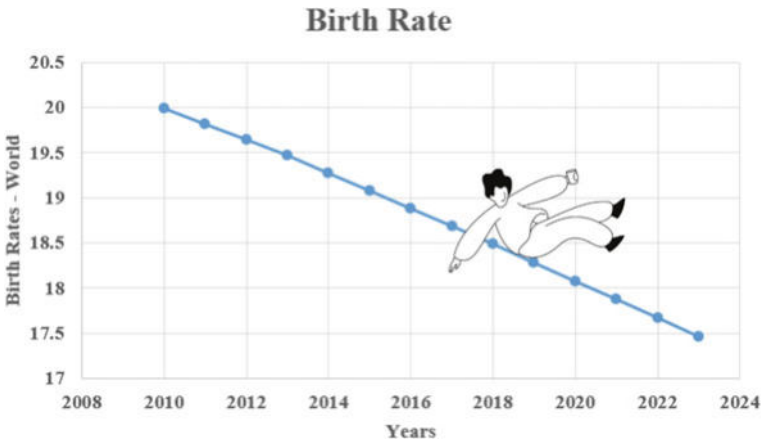


Fig. 4.6 Decline in birth rates

Factors Influencing Fertility Rates

Numerous factors influence fertility rates, and they can vary between countries and regions. Some of the key factors include:

1. Economic Development: As countries progress economically, fertility rates often decline. Industrialization and urbanization lead to increased education, employment opportunities, and access to family planning services, all of which contribute to lower fertility rates.
2. Education and Women's Empowerment: Women's education and empowerment play a significant role in shaping fertility rates. Educated women tend to marry later and have fewer children, as they prioritize career advancement and family planning.
3. Access to Family Planning Services: Availability and accessibility of family planning services and contraceptives influence family size decisions. Countries with well-established family planning programs typically experience lower fertility rates.
4. Cultural and Social Norms: Cultural and social factors, including traditional beliefs about family size and gender roles, can impact fertility rates. In some societies, large families are still preferred, while others may have strong support for smaller families.
5. Child Mortality and Healthcare: High child mortality rates can influence fertility decisions. In regions with higher child mortality, parents may choose to have more children as a form of social insurance against child loss.
6. Government Policies: Government policies related to family planning, parental leave, child benefits, and childcare can influence fertility rates. Pro-natalist policies may encourage higher birth rates, while pro-family policies might support work-life balance and women's employment.

Implications of Fertility Rates

1. Population Growth or Decline: Fertility rates directly impact population growth. A TFR above 2.1 leads to population growth, while a TFR below 2.1 may result in population decline over time.
2. Aging Population: Low fertility rates contribute to population aging. As birth rates decline and life expectancy increases, the proportion of elderly individuals in the population rises, posing challenges for pension systems and healthcare services.
3. Labor Force and Economic Impact: Fertility rates influence the size and composition of the labor force. A shrinking workforce due to low birth rates can impact economic productivity and sustainability of social welfare programs.
4. Dependency Ratio: Fertility rates affect the dependency ratio, which measures the number of non-working individuals (children and the elderly) supported by the working-age population. High fertility rates can lead to a larger dependent population.

The Measurement of Fertility Rates

Fertility rates are usually measured using three main indicators: the Total Fertility Rate (TFR) Age-Specific Fertility Rate (ASFR), and General Fertility rate (GFR).

1. Total Fertility Rate (TFR): The Total Fertility Rate represents the average number of children a woman would have if she were to pass through her reproductive years (ages 15–49) experiencing the prevailing age-specific fertility rates for each year. A TFR of 2.1 is considered the replacement level, meaning that, on average, each woman would give birth to enough children to replace herself and her partner in the population.
2. Age-Specific Fertility Rate (ASFR): The Age-Specific Fertility Rate refers to the number of live births per 1000 women in specific age groups (e.g., 15–19, 20–24, 25–29, etc.) during a given period, usually a year.
3. General Fertility Rate (GFR): The General Fertility Rate is a key demographic indicator that measures the number of live births per 1000 women of reproductive age (usually defined as women aged 15–49) in a given population during a specific period, typically a year.

The General Fertility Rate provides valuable information about the fertility patterns and reproductive behavior of women within a particular population. It is a more specific indicator compared to the Total Fertility Rate (TFR), which represents the average number of children a woman is expected to have during her reproductive years. It is commonly used in demographic studies and by policymakers to monitor and analyze changes in birth rates over time.

Calculating the General Fertility Rate involves dividing the number of live births in a specific time period by the number of women in the reproductive age group and

then multiplying the result by 1000 to express it as the number of births per 1000 women.

The formula for calculating the General Fertility Rate (GFR) is as follows:

GFR =(Number of Live Births in a Specific Time Period/Number of Women Aged 15–49) ×1000.

Example 1: If a country had 100,000 live births in a year and the number of women aged 15–49 in that country was 1,000,000, the General Fertility Rate would be:

GFR =(100,000/1,000,000) ×1000 =100 births per 1000 women aged 15–49.

TFR represents the average number of children a woman would give birth to during her reproductive years, assuming that the current age-specific fertility rates remain constant throughout her lifetime.

TFR =(Total number of live births)/(Number of women aged 15–49).

TFR =100,000/1000,000 TFR =0.1

The Total Fertility Rate for this country is 0.1, which means that, on average, each woman in the reproductive age range gives birth to 0.1 children.

ASFR is usually calculated for five-year age groups (e.g., 15–19, 20–24, 25–29, and so on) within the reproductive age range. Since we only have the total number of live births for the entire reproductive age range, we will not be able to calculate specific ASFR values in this case.

Example 2: Calculate the TFR, GFR, and ASFR for the following data (Tables 4.13 and 4.14).

Solution

ASFR in the above solution is computed using the formula,

$$\text{A.S.F.R} = \frac{\text{No of live births in a specific period of fertile period of women}}{\text{Women population in the age group}} * 100$$

The GFR is given as,

$$\text{G.F.R} = \frac{\text{No of live births occuring in a year}}{\text{Average population of women of child bearing age}} * 100 \quad \frac{44{,}888}{532{,}660} = \mathbf{8.42}$$

Table 4.13 Data to calculate the TFR, GFR, and ASFR

Age (years)	Women population	Number of births to women
15–19	88,790	345
20–24	70,910	14,544
25–29	72,660	16,740
30–34	78,920	12,267
35–39	75,100	876
40–44	79,620	56
45–49	66,660	60

Table 4.14 Solution for calculating the TFR, ASFR, and GFR for the data

Age (years)	Women population	Number of births to women	ASFR
15–19	88,790	345	3.89
20–24	70,910	14,544	205.11
25–29	72,660	16,740	230.39
30–34	78,920	12,267	155.44
35–39	75,100	876	11.66
40–44	79,620	56	0.70
45–49	66,660	60	0.90
Total	532,660	44,888	608.083

Net Production Rate

$$\text{T.F.R} = 5 * \sum \text{A.S.F.R} = 5 * 608.0825 = \mathbf{3040.413}$$

This refers to a demographic measure used to estimate population growth. It is also known as the net reproduction rate or the net reproductive rate. The net production rate specifically focuses on the female population and measures the average number of daughters that would be born to a woman during her lifetime, given the current age-specific fertility rates.

Here is how the net production rate is calculated:

1. Age-Specific Fertility Rates: The first step is to determine the age-specific fertility rates, which represent the number of live births occurring to women in specific age groups. These rates are usually calculated per 1000 women of that age group.
2. Cohort Component Method: The net production rate uses a cohort-component method to estimate the average number of daughters born to a woman over her reproductive lifetime. This method involves tracking a hypothetical cohort of females throughout their reproductive ages (usually from 15 to 49 or 15 to 50).
3. Summation: For each age group, the number of daughters expected to be born to each woman in the cohort is multiplied by the corresponding age-specific fertility rate. The products are then summed across all age groups to obtain the net production rate.

Interpretation:

- If the net production rate is greater than 1, it indicates that each generation of women is producing more daughters than themselves, leading to population growth.
- If the net production rate is equal to 1, it suggests that each generation of women is replacing themselves, and the population is stable (zero population growth).
- If the net production rate is less than 1, it means that each generation of women is producing fewer daughters than themselves, leading to population decline.

The net production rate is a crucial tool in understanding population dynamics, fertility patterns, and long-term population projections. It helps demographers and policymakers make informed decisions about issues such as family planning, healthcare, and social services.

Example 3: Calculate the net production rate from the following data (Table 4.15):

Solution

$$\text{N.P.R} = \text{i} * \sum \text{S} * \text{W.S.F.R} \ = 5 * 221.97 = \mathbf{1109.86},$$

where *i* refers to the class interval.

$$\text{N.P.R. per women} \ = \mathbf{1.109}$$

Conclusion: Since the N.P.R is greater than 1, the population is said to be on a rise.

Net production rates, also known as population growth rates or natural increase rates, are important statistics in demography that provide insights into the dynamics of a population. These rates are calculated by comparing the number of births and deaths within a population over a specific period. Here are some interesting facts about net production rates and related vital statistics:

1. Net Production Rate Calculation: The net production rate is typically calculated as the difference between the crude birth rate (CBR) and the Crude Death Rate (CDR) of a population. It is expressed as a percentage or a decimal. The formula is as follows:

$$\text{Net Production Rate} \ = \ (\text{Crude Birth Rate} \ - \text{Crude Death Rate}) \ \times 100$$

2. Positive and Negative Rates: When the crude birth rate exceeds the Crude Death Rate, the net production rate is positive, indicating population growth.

Table 4.15 Women population, survival rates, and number of female births to women, categorized based on their age

Age (years)	Women population	No of births to women	W.S.F.R	Survival rate (S)	S*W.S.F.R
15–19	4806	102	21.2	0.9540	20.25
20–24	5423	432	79.7	0.9470	75.44
25–29	5434	318	58.5	0.9370	54.83
30–34	3814	168	44.0	0.9290	40.92
35–39	3567	76	21.3	0.9170	19.54
40–44	3221	27	8.4	0.9050	7.59
45–49	2611	10	3.8	0.89	3.41
Total		1133	236.97	6	221.97

Conversely, if the Crude Death Rate surpasses the crude birth rate, the net production rate is negative, indicating population decline.

3. Replacement-Level Fertility: The net production rate required for a population to replace itself without migration is called the replacement-level fertility. It is typically slightly above 2.0, accounting for mortality and childlessness. If the net production rate is above the replacement level, the population is growing; if it is below, the population is declining.
4. Demographic Transition: The net production rate is closely linked to the demographic transition theory. As societies undergo economic and social development, they typically experience a transition from high birth and death rates to low birth and death rates. This transition leads to population growth in the early stages and eventual stabilization in later stages.
5. Impact of Net Production Rates on Age Structure: High net production rates contribute to a youthful population with a large proportion of children and young adults. Conversely, low or negative net production rates result in an aging population with a higher proportion of older individuals.
6. Regional Variations: Net production rates vary significantly across countries and regions due to differences in socioeconomic factors, cultural norms, access to healthcare, and family planning practices. Developing countries generally have higher net production rates compared to developed countries.

4.9 Marriage and Divorce Statistics

Marriage and divorce statistics vary across different countries and regions, but some common trends observed globally are as follows:

Marriage Statistics:

1. Marriage Rates: The number of marriages per 1000 people in a given population.
2. Average Age at Marriage: The average age at which people get married. This has generally been increasing in many developed countries, with more individuals choosing to marry later in life.
3. Marriage Duration: The average length of marriages before divorce or separation.
4. Marital Status: The percentage of the population that is married, divorced, widowed, or never married.

Divorce Statistics:

1. Divorce Rates: The number of divorces per 1000 married individuals in a given population.
2. Average Age at Divorce: The average age at which people get divorced.
3. Divorce Reasons: The most common reasons cited for divorce, such as infidelity, financial issues, and communication problems.
4. Divorce Duration: The average time it takes for a divorce to be finalized.

Fig. 4.7 Marriage and divorce rates in the United States (per 1000). *Source* https://blogs.sas.com/content/sastraining/2015/08/04/marriage-and-divorce-in-the-us-what-do-the-numbers-say/

5. Divorce Rates Among Different Demographics: How divorce rates vary among different age groups, socioeconomic backgrounds, and cultural or religious affiliations.

It is important to note that while these statistics provide insights into marriage and divorce trends, they can vary greatly depending on the cultural, societal, and legal norms of each country or region (Fig. 4.7).

Interesting Divorce Facts

- In the United States, there is one divorce approximately every 36 s. That is nearly 2400 divorces per day, 16,800 divorces per week, and 876,000 divorces a year [iii].
- The average length of a first marriage that ends in divorce is 8 years [iv].
- The probability of a first marriage ending in separation or divorce in the first 5 years is 20% and in 10 years is 33% [v].
- The average age for couples going through divorce is 30 years old.
- On average, a person spends about two years thinking about divorce before taking action.
- About 3 out of every 4 divorced people will remarry [vi].
- People wait an average of 3 years after a divorce to remarry (if they remarry at all).
- Six percent of divorced couples end up remarrying each other [vii].

Fig. 4.8 Cost of divorce https://financesonline.com/divorce-statistics/

- The US government stopped collecting detailed marriage and divorce statistics in 1996, so other data sources, such as the US Census and independent researchers, are used to estimate divorce rates and other statistics (Fig. 4.8).

https://www.mckinleyirvin.com/family-law-blog/2012/october/32-shocking-divorce-statistics/

Many anticipated that the COVID-19 pandemic would be problematic for married couples. After all, 4 in 10 adults in the United States reported signs of mental health issues during the pandemic (Panchal et al., 2021). Surprisingly, divorces dropped across nearly all United States. A survey even shows that couples deepened their bonds while on lockdown.

https://financesonline.com/divorce-statistics/#2

- The number of troubled marriages dropped from 40% in 2019 to 29% in 2020 during the pandemic (Wang, 2020).
- Similarly, the number of divorces in Florida declined by 28% (Steverman, 2021).
- But the biggest decline in divorces recorded among American States occurred in New Hampshire, with a decrease of 36.4% (Manning & Payne, 2020).
- Conversely, divorces in Arizona increased by 9% (Wang, 2020).
- 58% of married Americans admit the COVID-19 lockdown made them value their partner more (Wang, 2020).
- With *many employees working from home* due to the pandemic, 50% of married US couples were able to spend more time with each other and have deepened their relationship (Wang, 2020).
- 44% of people who increased their income during the pandemic revealed that COVID-19 did not cause stress to their marriage (Wang, 2020).
- However, 45% of couples whose incomes declined state that the pandemic caused stress to their marriage (Wang, 2020).
- Unfortunately, reports of domestic violence increased by 27% in Jefferson Country, Alabama, 22% in Portland, and 18% in San Antonio, Texas (Boserup, 2020).

- In Saudi Arabia, the divorce rate increased by 30% during the lockdown period (Barakat, 2020).
- And in the same way, the number of divorces in China surged by 10–20% (Chamadia, 2020).

4.10 Life Tables

Introduction: The average additional years that a person could anticipate to live if present mortality trends persisted for the remaining years of their lives is known as life expectancy. According to the current age-specific death rates, a life table is a tabular representation of life expectancy and the likelihood of dying at each age or age group for a given population.

The life table provides an organized, comprehensive view of the mortality in a population.

4.10.1 Why Do We Need Life Tables?

- On the basis of the current death rate, a life table is used to predict the population for the future.
- Based on age-specific death rates, it aids in calculating the average life expectancy.
- To determine the causes of certain death rates, male and female death rates, etc., the method of creating a life table can be used.
- The net migration rate can be calculated using the survival rates in a life table based on age distribution at 5- or 10-year intervals.
- Population trends at the local, national, and international levels can be compared using life tables.
- Marriage trends and variations in them can be estimated by building a life table depending on the age of the marriage.
- For the analysis of socioeconomic data in a nation, several decrement life tables relating to cause-specific death rates, male and female death rates, etc. can be produced in place of a single life table.
- The creation of family planning programs that address infant mortality, maternal mortality, health programs, etc. makes use of life tables in particular. They may also be employed to assess family planning initiatives.
- Life insurance firms now utilize life tables to assess the average life expectancy of individuals, separately for males and females. They assist in calculating the premium that a person in a given age group must pay.
- Additionally, the life table gives the insurance company financial support so it will not suffer a loss and enables it to pay the insured sum to the decedent's legitimate heirs in the event of the insured person's death before the policy's maturity.

4.10.2 Examples Where Life Tables Are Used

A life table is a table-based summary of a community's mortality experiences over a specific time period. The life table displays the number of people living and dying at each age based on a cohort's experience. It also indicates the likelihood of dying and living separately. The life table depicts a cohort's life history.

- A cohort is a group of people who were born at the same time and died under similar circumstances.
- Actuaries use the life table to calculate premium rates for people of various ages.
- It aids in determining the accuracy of census figures, death and birth registrations.
- It aids in assessing the impact of family planning on population growth.
- It allows us to assess the increase in life span as a result of new scientific discoveries, sophisticated medical treatments, and better living conditions.
- Migration estimates can be derived from the life table.

4.10.3 Other Applications of Life Tables

Aside from its use in insurance life tables, it can be used to conduct comparative analyses of mortality conditions across countries or regions. The following are some of the applications of life tables:

1. Calculation of mortality due to specific causes: For comparisons, life tables are calculated for different groups of the population such as sex (male/female), age distribution (different age groups), religion, and region. The mortality statistics may prompt us to investigate the specific causes of death in various groups of people.
2. Comparison of mortality conditions: The best indices of mortality are life expectancy at birth and other ages. These indices vary greatly from place to place and over time. Because of improved health care, life expectancy in most countries has steadily increased over time.
3. Population projections: The life tables were also used to create population projections by age and gender. This is useful in estimating the population size at some future point.

4.10.4 Limitations of Life Tables

Life tables are created using demographic data from sources such as the census and SRS. As a result, life table estimates suffer from all of the drawbacks of a statistical measure based on population censuses and vital records. Age and mortality registration data may be incomplete or biased. Because infant mortality has a significant impact on life expectancy, under-reporting of this indicator, which is common in many countries, can have a significant impact on the results of the tables.

Furthermore, significant differences in specific age groups with high mortality may be overlooked because they have little effect on overall life expectancy.

It is generally not recommended to create life tables for small populations at the local or subregional level because migratory movements affect population structure more than at the regional or national levels. In these cases, a very small number of deaths can be obtained, resulting in inaccurate calculations of the table's columns.

4.11 Life Tables—Basic Notations

See Table 4.16.

4.11.1 Life Expectancy

Life expectancy is a critical measure in demographic analysis. It represents the average number of years a person is expected to live from a given age under the assumption that current age-specific mortality rates remain constant throughout their lifetime. Life expectancy is often reported at birth, but it can also be calculated for any age group. The indicator is a powerful reflection of the overall health and living conditions within a population.

Table 4.16 Life table notations

S. No.	Notation	Definition
1	x, x +n	Age interval or period of life between two exact ages stated in years
2	${}_nq_x$	Proportion of persons alive at the beginning of the age interval who die during the age interval
3	l_x	Of the starting number of newborns in the life table (called the radix of the life table, usually set at 100,000) the number living at the beginning of the age interval (or the number surviving to the beginning of the age interval)
4	${}_nd_x$	The number of persons in the cohort who die in the age interval (x, x +n)
5	${}_nL_x$	Number of years of life lived by the cohort within the indicated age interval (x, x +n) (or person-years of life in the age interval)
6	T_x	Total person-years of life contributed by the cohort after attaining age x
7	e_x^0	The average number of years of life remaining for a person alive at the beginning of age interval x

4.11.2 Abridged Life Table

Abridged life tables are invaluable tools for understanding population health and mortality patterns. Their simplicity and efficiency make them widely used in demography and vital statistics. By providing critical information about life expectancy and mortality rates across age groups, abridged life tables assist policymakers, researchers, and public health professionals in making informed decisions to improve the well-being of populations around the world.

While traditional life tables provide comprehensive insights into population mortality, they require extensive data and substantial computational efforts, making them resource-intensive and time-consuming to construct. To address this challenge, demographers developed abridged life tables as a more practical alternative.

Abridged life tables offer a simplified version of the standard life table. Instead of reporting mortality rates for each age, they group ages into intervals, such as five-year or ten-year age groups. By using these age intervals, abridged life tables significantly reduce the amount of data needed while still providing reasonably accurate estimates of life expectancy and other mortality indicators.

4.11.3 Construction of Abridged Life Tables

The construction of abridged life tables involves several steps:

1. Age Grouping: The first step is to group the population into appropriate age intervals. The choice of intervals depends on the available data and the specific research or policy objectives.
2. Exposure and Deaths: The number of individuals exposed to the risk of dying (i.e., the population within each age group) and the number of deaths that occurred in each age group during the observation period are recorded.
3. Probability of Dying: The probability of dying, often referred to as the age-specific mortality rate, is calculated for each age interval. This is done by dividing the number of deaths in an age group by the population within that age group.
4. Life Expectancy: Using the probability of dying, one can calculate the life expectancy for each age interval. These life expectancies are then used to calculate the overall life expectancy for the entire population.

4.11.4 Significance of Abridged Life Tables

Abridged life tables are an indispensable tool in population health analysis due to their numerous advantages:

1. Resource Efficiency: Abridged life tables require less data and computation compared to standard life tables, making them more accessible and quicker to construct.
2. Policy Insights: Governments and public health authorities use abridged life tables to gain insights into the mortality patterns of specific age groups. This information aids in the design and evaluation of targeted health interventions.
3. International Comparisons: Abridged life tables allow for easy comparison of mortality patterns between different countries or regions, enabling researchers to identify disparities and trends.
4. Projection Purposes: These life tables serve as a foundation for projecting future mortality trends, assisting in long-term planning for healthcare and pension systems.

4.11.5 Limitations

Despite their advantages, abridged life tables do have some limitations. The grouping of ages may lead to a loss of granularity, making it difficult to capture subtle variations in mortality within specific age groups. Additionally, because they assume that mortality rates remain constant within each interval, they may not account for certain factors that influence mortality changes over time.

Example: In a city of Bangladesh, the people were exposed to a certain airborne disease that arouse due to a typhoon in the region. This typhoon has a few suspended harmful particles in the air which has led to many lung congestion and breathing issues in people aged above 40 years. Starting with 10,000 persons of age 40, construct a mortality table up to the age 50, using the following values of q_x from the mortality rates. The medical researcher wants to understand the death patterns in the city also predict the probability that a person aged 40 will live for 5 more years with good health conditions (Tables 4.17 and 4.18).

Table 4.17 Effect of epidemics in Bangladesh

Age (x)	40	41	42	43	44	45
Mortality rate (q_x)	0.0122	0.0135	0.0138	0.0145	0.0148	0.0151
Age (x)	**46**	**47**	**48**	**49**	**50**	
Mortality rate (qx)	0.0165	0.017	0.0178	0.0186	0.0198	

Table 4.18 Solution table

Age (x)	q_x	p_x	l_x	d_x	$_nL_x$	T_x	$e_x{}^0$
40	0.0122	0.98780	10,000	122	9939	1,01,634	10.16343
41	0.0135	0.98650	9878	133	9811	91,695	9.282777
42	0.0138	0.98620	9745	134	9677	81,884	8.402967
43	0.0145	0.98550	9610	139	9540	72,207	7.513554
44	0.0148	0.98520	9471	140	9401	62,666	6.616747
45	0.0151	0.98490	9331	141	9260	53,265	5.708635
46	0.0165	0.98350	9190	152	9114	44,005	4.788491
47	0.017	0.98300	9038	154	8961	34,891	3.8 60438
48	0.0178	0.98220	8884	158	8805	25,930	2.918554
49	0.0186	0.98140	8726	162	8645	17,124	1.962384
50	0.0198	0.98020	8564	170	8479	8479	0.9901
Total			**8394**				

Solution

$$\text{The probability that t the person aged 40 will live for 5 more yars } = \frac{l_{45}}{l_{40}} = \frac{9331}{10000} = \mathbf{0.9331}$$

4.12 Case Studies Related to Vital Statistics

Infant Mortality Rate Reduction:

In a particular country, the government implemented various interventions to address the high infant mortality rate (IMR). These interventions included improving access to prenatal care, promoting breastfeeding, enhancing healthcare infrastructure, and strengthening immunization programs. Over a five-year period, the IMR decreased significantly from 45 deaths per 1000 live births to 25 deaths per 1000 live births. This case study highlights the effectiveness of targeted interventions in reducing infant mortality.

Aging Population and Life Expectancy:

A country experienced a significant increase in life expectancy due to improved healthcare and socioeconomic conditions. As a result, the proportion of the population aged 65 and above increased substantially. This demographic shift posed challenges for the healthcare system, retirement planning, and social support structures. The government responded by implementing policies to accommodate the

needs of the aging population, including healthcare reforms, long-term care facilities, and pension reforms. This case study demonstrates the implications of an aging population on vital statistics and the need for appropriate policy responses.

Maternal Health and Maternal Mortality Reduction:

A region with high Maternal Mortality Rates implemented a comprehensive maternal health program. This program focused on improving access to skilled birth attendants, ensuring emergency obstetric care, promoting family planning, and educating communities about maternal health issues. As a result, Maternal Mortality Rate decreased significantly over a period of five years. This case study emphasizes the importance of addressing maternal health to reduce maternal mortality and improve vital statistics related to maternal and child health.

Epidemiological Investigation of Disease Outbreak:

During an outbreak of a novel infectious disease, epidemiologists conducted a detailed investigation to understand the patterns and impact of the disease. They collected data on the number of cases, age distribution, geographic distribution, and mortality rates. Through rigorous analysis, they identified risk factors, transmission pathways, and effective control measures. This case study illustrates how vital statistics, along with epidemiological investigations, play a crucial role in understanding and managing disease outbreaks.

Population Growth and Urbanization:

In a developing country, rapid urbanization led to a significant increase in population size and density in urban areas. This population growth posed challenges in providing adequate housing, healthcare, education, and infrastructure. The government implemented urban planning strategies, such as building affordable housing, improving transportation networks, and expanding healthcare facilities, to address these challenges. This case study highlights the interplay between population growth, urbanization, and the need for well-planned infrastructure and services to sustain healthy and thriving communities.