

# Business Analytics

The Science of Data-Driven Decision Making

U Dinesh Kumar



WILEY

# Business Analytics

The Science of Data-Driven Decision Making



# Business Analytics

The Science of Data-Driven Decision Making

**U Dinesh Kumar**

*Professor, Decision Sciences and Information Systems  
Indian Institute of Management Bangalore*

**WILEY**

# Business Analytics

## The Science of Data-Driven Decision Making

Copyright © 2017 by Wiley India Pvt. Ltd., 4435-36/7, Ansari Road, Daryaganj, New Delhi-110002.

Cover Image: Ali Kerem Yücel/Getty Images

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or scanning without the written permission of the publisher.

**Limits of Liability:** While the publisher and the author have used their best efforts in preparing this book, Wiley and the author make no representation or warranties with respect to the accuracy or completeness of the contents of this book, and specifically disclaim any implied warranties of merchantability or fitness for any particular purpose. There are no warranties which extend beyond the descriptions contained in this paragraph. No warranty may be created or extended by sales representatives or written sales materials. The accuracy and completeness of the information provided herein and the opinions stated herein are not guaranteed or warranted to produce any particular results, and the advice and strategies contained herein may not be suitable for every individual. Neither Wiley India nor the author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

**Disclaimer:** The contents of this book have been checked for accuracy. Since deviations cannot be precluded entirely, Wiley or its author cannot guarantee full agreement. As the book is intended for educational purpose, Wiley or its author shall not be responsible for any errors, omissions or damages arising out of the use of the information contained in the book. This publication is designed to provide accurate and authoritative information with regard to the subject matter covered. It is sold on the understanding that the Publisher is not engaged in rendering professional services.

**Trademarks:** All brand names and product names used in this book are trademarks, registered trademarks, or trade names of their respective holders. Wiley is not associated with any product or vendor mentioned in this book.

Other Wiley Editorial Offices:

John Wiley & Sons, Inc. 111 River Street, Hoboken, NJ 07030, USA

Wiley-VCH Verlag GmbH, Pappellaee 3, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 42 McDougall Street, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada, M9W 1L1

First Edition: 2017

ISBN: 978-81-265-6877-2

ISBN: 978-81-265-8384-3 (ebk)

[www.wileyindia.com](http://www.wileyindia.com)

*To Haritha and Prathusha*

*To the one who showed a lotus flower when asked to teach*



# Preface

When the student is ready, the teacher will appear.

— The Buddha

Business Analytics has become one of the most important skills that every business school student should acquire to become successful in a management career. The use of analytics across industries for decision making, automation of business processes, products, and solutions driven by analytics makes it an essential skill for every student graduating from management and engineering disciplines. Analytics is used as a competitive strategy by many successful companies. Many organizations generate solutions to their problems using analytics and innovation in many companies is driven by analytics. Organizations such as Amazon, Apple, Bigbasket, Capital One, Disneyland, Facebook, Flipkart, Google, IBM, and Walmart have created solutions using analytics. Amazon has created solutions such as recommender systems and Amazon Go that are driven by analytics. Apple's predictive keyboard is another example of solutions driven by analytics. Analytics is relevant not only to profit-making companies but also for government and non-government organizations (NGOs). The Akshaya Patra Foundation, an NGO based out of Bangalore, has used several analytics models for effective management of its free meal program which provides free meals to about 1.5 million school children in India. Governments across the world use analytics to handle black money and other economic offences.

One of the challenges that organizations face in analytics is finding the right talent. There is a severe shortage of analytics talent across the globe. International Data Corporation (IDC) predicted a shortage of 1,81,000 data scientists by 2018 and a severe shortage was also reported by Deloitte and McKinsey. It is clear from the reported literature that there is significant shortage of talent in data science. Many universities and institutions have started several undergraduate and post-graduate programs in Business Analytics. Having said that, there is shortage of quality teaching material in Business Analytics. This book makes an attempt to fill the gap. There are several books on analytics written for practitioners that focus on strategic and tactical applications of analytics. There are also many books that explore the theoretical aspects of analytics. I have chosen the middle path, discussing the theoretical foundation of analytics as well as its applications and deployment using case studies. The objective of the book is to introduce the students to foundations of data science and expose them to several advanced analytics techniques using several applications and case studies from Indian organizations.

I have been teaching Business Analytics at the Indian Institute of Management Bangalore since 2008 as an elective course for the post graduate program in management. I started a certificate program in analytics at IIMB in 2010 and have conducted several training programs and provided consulting services in Analytics. The book is a collection of material created over the last 10 years and was tested among students of post graduate program in management and working executives across several corporates.

## ORGANIZATION OF THE BOOK

The book has 17 chapters and addresses all components of analytics such as descriptive, predictive, and prescriptive analytics. The first few chapters are dedicated to foundations of business analytics. Introduction to business analytics and its components such as descriptive, predictive, and prescriptive analytics along with several applications are discussed in Chapter 1. In Chapters 2 to 8, we discuss basic statistical concepts such as descriptive statistics, concept of random variables, discrete and continuous random variables, confidence interval, hypothesis testing, analysis of variance and correlation. Chapters 9 to 13 are dedicated to predictive analytics techniques such as multiple linear regression, logistic regression, decision tree learning, and forecasting techniques. Clustering is discussed in Chapter 14. Chapter 15 is dedicated to prescriptive analytics in which concepts such as linear programming, integer programming, and goal programming are discussed. Stochastic models and Six Sigma are discussed in Chapters 16 and 17, respectively. Each chapter is enriched with several solved problems, multiple choice questions, and exercise problems. Many chapters include case studies that will improve understanding of the concepts and enable students to understand how analytics is used by the industry.

The book is written in such a way that a beginner can learn the concepts with little or no assistance from the instructor. Each chapter starts with learning objectives and ends with exercise problems. Students without any statistical background should go through the initial chapters before reading predictive analytics chapters. The multiple choice questions at the end of the chapter will ensure that the reader has understood the learning outcomes. The best way to understand the concepts is to solve the exercise problems and case studies. The microsite of the book lists the data sets used in the book. The case studies in this book are distributed through the Harvard Business Publishing (HBP) case portal. Several institutions across the world have been using case studies for teaching analytics.

## FEATURES OF THE BOOK

Several features are incorporated in the book to make the learning effective and fun.

### Learning Objectives

Learning objectives provide a summary of key concepts that would be discussed in the chapter and key takeaways from the chapter.



#### LEARNING OBJECTIVES

- LO 8-1** Understand the concept of correlation and its role in analytics.
- LO 8-2** Learn to calculate correlation between two continuous variables.
- LO 8-3** Understand the difference between correlation and causation.
- LO 8-4** Understand correlation between a continuous variable and a discrete variable.
- LO 8-5** Learn to calculate correlation between two discrete variables.

## Fun Facts

The book is filled with several fun facts about insights derived from both correct and incorrect use of data.



### 4.1.1 | When Jesus Christ Became 2<sup>nd</sup> Best in the World

Respondent bias is another source problem in survey sampling. There was an internet poll conducted in 1998 to find the 'most influential figure of the last 2000 years'. Jamie Pollock who played as defensive mid-fielder for the football team Manchester City won the title leaving Jesus Christ and Carl Marx to second and third place, respectively (Szczepanik, 2016). Apparently the poll was rigged by the supporters of the club Queens Park Rangers (QPR) who voted multiple times for Jamie Pollock. The story goes like this: On 25<sup>th</sup> April 1998, Manchester City was playing against QPR in Division 1 of English football and both teams were in danger of being relegated to division 2 (third tier of English football). When the goal was 1–1, Jamie Pollock scored an own goal giving upper hand to QPR. Although the match ended in a 2–2 draw, Manchester City was relegated to Division 2 for the first time and QPR managed stay in Division 1 (Moxley 2012, Szczepanik 2016). Collecting survey and ensuring that the sample is unbiased is one of the major challenges in analytics. Summers (1969) grouped bias in survey research in to several categories such as (a) sampling bias, (b) non-responsive bias, (c) respondent bias, (d) instrumentation bias, etc.

## Case Study

Best way to learn analytics is to understand the business context, deriving analytics problem from business context and using data science to derive insights and actionable items. The book has 9 case studies on use of analytics by Indian companies.



### Case Study

#### Customer Analytics at Flipkart.com

It was typical cloudy monsoon weather at Bangalore on July 28, 2015. In the Darwin room of Flipkart's Cessna Business Park office, Ravi Vijayaraghavan, the Head of Analytics and Pravin Shinde, Senior Manager Analytics were brainstorming various business problems that Flipkart as an e-commerce company was encountering. Flipkart had been putting in much effort and emphasis on the use of analytics in every aspect of decision making. Forecasting demand for thousands of stock-keeping units (SKUs), predicting returns and cancellations of orders, predicting the reasons when customers contact the customer service centers, optimizing markdown pricing, identifying various types of frauds, optimizing vehicle routing, and enabling adherence to service-level agreements, were some of the typical problems that the analytics division of Flipkart was solving using state-of-the-art analytics techniques. In 2015, the team included about 100 data scientists mostly recruited from institutes such as the Indian Institute of Technology and Indian Institute of Management specifically for this purpose.

## Supplementary Online Material

The readers can enhance their learning experience by watching the predictive analytics course offered by the author on edX platform.

The course is part of the Indian Institute of Management Bangalore's massive open online course (MOOC). The course videos are available at the following link:



<https://www.edx.org/course/predictive-analytics-iimbx-qm901x>

## Multiple Choice Questions

Multiple choice questions ensure that the students have understood the important concepts of the chapter.



### MULTIPLE CHOICE QUESTIONS

1. Estimated value of parameter from a sample is called
 

(a) Sample statistic	(b) Population parameter	(c) Unbiased estimator
(d) Sampling		
2. Sampling frame is used for
 

(a) Calculating the sample size	(b) Identifying the source of data
(c) Estimating population parameter	(d) Estimating statistic

### EXERCISES

1. On 8 November 2016, the Indian government demonetized 500 and 1000 rupees notes and allowed the citizens to deposit the old notes in the banks. The average amount deposited by a customer at a bank in Bannerghatta road is INR 17500 and the corresponding standard deviation is 4500. At a bank on Bannerghatta Road, Bangalore, 156 customers deposited money on 11 November 2016.

(a) Calculate the probability that the total deposits exceed INR 3 million.



## Exercise Problems

Exercise problems help the students to understand the concepts using multiple scenarios and have an inherent objective of strengthening the conceptual understanding.



## Data Set

An important resource for learning analytics is the data sets. The data sets used in the book can be downloaded from the following website.

<https://www.wileyindia.com/business-analytics-the-science-of-data-driven-decision-making.html>

# Acknowledgments

As a faculty member at IIMB, I have the privilege of using all the facilities at IIMB and an opportunity to interact with other intellectuals both from industry and academia. I would like to thank IIMB for giving permission to reproduce the following case studies<sup>1</sup>.

1. Central Parking Services Private Limited (IMB 451)
2. Pricing of Players in the Indian Premier League (IMB 379)
3. HR Analytics at Scaleneworks – Behavioural Model to Predict Reneging (IMB 551)
4. Breaking Barriers – Micro Mortgage Analytics (IMB 446)
5. Markdown Optimization for an Indian Apparel Retailer (IMB 561)
6. Larsen and Toubro: Spare Parts Forecasting (IMB 499)
7. Managing Linen at Apollo Hospitals (IMB 495)
8. Customer Analytics at Flipkart.com (IMB 555)
9. Era of Quality at the Akshaya Patra Foundation (IMB 493)

The case studies discussed in the book are not intended to serve as an endorsement, source of primary data or to show effective or inefficient handling of decision or business processes. I would like to thank IIMB for providing all the facilities and giving me an opportunity to start Data Centre and Analytics Lab (DCAL) as a center of excellence. Activities of DCAL have contributed immensely towards development of this book.

I would like to thank the Akshaya Patra Foundation, Apollo Hospitals, Central Parking Solutions, Larsen and Toubro, Flipkart, Scaleneworks, and Shubham Housing Finance Limited for providing the data and permitting us to write case studies.

A textbook of this size would not have been possible without the assistance of my colleagues, friends, and students. I would like to thank my colleagues at IIMB for their support during the development of this book. I thank all my case co-authors Abhishek Srivastava, Apoorva Sara Prakash, Dhimant Ganatra, Kshitiz Ranjan, Muthu Solayappan, Prakash Hegde, Naveen Bhansali, Jitendra Rudravaram, Rahul Kumar, Ruchi Jaiswal, Shailaja Grover, Suhruta Kulkarni, Srujana, and Tanmay Gupta for assisting me to write the cases. Special thanks to Sheetal Malagi who read through all the chapters of the book and provided critical reviews on each chapter and collected supporting material. I gratefully acknowledge the assistance received from Aarti Krishnan, Arun Pandit, Bharath Paturi, Dhimant Ganatra, Manaranjan Pradhan, Purvi Tiwari, Sandhya Shenoy, Satyabala Hariharan, Shailaja Grover, and Sunil throughout the preparation of this book. I was lucky to have good support staff to help me whenever

<sup>1</sup> Cases of the Indian Institute of Management Bangalore are prepared as a basis of class discussion. Cases are not designed to present illustrations of either correct or incorrect handling of administrative problems.

needed. Chitralekha helped me with editing all the case studies, Nirmala and Sumithra provided all the administrative support.

I would like to thank Debarati Sengupta of Wiley for her support throughout the preparation of this book. Special thanks to Meenakshi Sehrawat of Wiley for carefully editing the book and offering several suggestions to improve the quality of the book.

**Dinesh Kumar**  
**Professor, Decision Sciences and Information Systems**  
**Indian Institute of Management, Bangalore**

# Contents

<i>Preface</i>	vii		
<i>Acknowledgments</i>	xi		
<b>1. Introduction to Business Analytics</b>	<b>1</b>		
1.1 Introduction to Business Analytics	2	2.3 Types of Data Measurement Scales	34
1.2 Why Analytics	4	2.3.1 Nominal Scale (Qualitative Data)	34
1.3 Business Analytics: The Science of Data-Driven Decision Making	7	2.3.2 Ordinal Scale	34
1.3.1 Business Context	8	2.3.3 Interval Scale	35
1.3.2 Technology	9	2.3.4 Ratio Scale	35
1.3.3 Data Science	9	2.4 Population and Sample	35
1.4 Descriptive Analytics	10	2.5 Measures of Central Tendency	35
1.5 Predictive Analytics	13	2.5.1 Mean (Or Average) Value	36
1.6 Prescriptive Analytics	15	2.5.2 Median (Or Mid) Value	37
1.7 Descriptive, Predictive, and Prescriptive Analytics Techniques	17	2.5.3 Mode	38
1.8 Big Data Analytics	18	2.6 Percentile, Decile, and Quartile	38
1.9 Web and Social Media Analytics	19	2.7 Measures of Variation	40
1.10 Machine Learning Algorithms	21	2.7.1 Range	40
1.11 Framework for Data-Driven Decision Making	22	2.7.2 Inter-Quartile Distance (IQR)	40
1.12 Analytics Capability Building	22	2.7.3 Variance and Standard Deviation	40
1.13 Roadmap for Analytics Capability Building	24	2.7.4 Chebyshev's Theorem	43
1.14 Challenges in Data-Driven Decision Making and Future	25	2.8 Measures of Shape – Skewness and Kurtosis	43
1.15 Organization of the Book	27	2.9 Data Visualization	45
<i>References</i>	27	2.9.1 Histogram	45
<b>2. Descriptive Analytics</b>	<b>31</b>	2.9.2 Bar Chart	48
2.1 Introduction to Descriptive Analytics	32	2.9.3 Pie Chart	49
2.2 Data Types and Scales	32	2.9.4 Scatter Plot	49
2.2.1 Structured and Unstructured Data	32	2.9.5 Coxcomb Chart	50
2.2.2 Cross-Sectional, Time Series, and Panel Data	34	2.9.6 Box Plot (or Box and Whisker Plot)	51
		2.9.7 Treemap	51
		<i>Summary</i>	53
		<i>Multiple Choice Questions</i>	53
		<i>Exercises</i>	54
		<i>References</i>	55
<b>3. Introduction to Probability</b>	<b>57</b>		
		3.1 Introduction to Probability Theory	57
		3.2 Probability Theory – Terminology	58

3.2.1	Random Experiment	58	3.8.4	Approximation of Binomial Distribution using Normal Distribution	75
3.2.2	Sample Space	58	3.9	Poisson Distribution	77
3.2.3	Event	59	3.10	Geometric Distribution	79
3.2.4	Probability Estimation Using Relative Frequency	59	3.10.1	Memoryless Property of Geometric Distribution	80
3.2.5	Algebra of Events	60	3.11	Parameters of Continuous Distributions	81
3.3	Fundamental Concepts in Probability – Axioms of Probability	61	3.12	Uniform Distribution	82
3.3.1	Joint Probability	61	3.13	Exponential Distribution	82
3.3.2	Marginal Probability	63	3.13.1	Memoryless Property of Exponential Distribution	83
3.3.3	Independent Events	63	3.14	Normal Distribution	85
3.3.4	Conditional Probability	63	3.14.1	Properties of Normal Distribution	86
3.4	Application of Simple Probability Rules – Association Rule Learning	64	3.14.2	Standard Normal Variable	87
3.4.1	Association Rule Learning	64	3.15	Chi-Square Distribution	90
3.5	Bayes' Theorem	66	3.15.1	Properties of Chi-Square Distribution	92
3.5.1	Solving Monty Hall Problem Using Bayes' Theorem	66	3.16	Student's <i>t</i> -Distribution	92
3.5.2	Generalization of Bayes' Theorem	67	3.16.1	Properties of <i>t</i> -Distribution	94
3.6	Random Variables	68	3.17	<i>F</i> -Distribution	94
3.6.1	Discrete Random Variables	69	3.17.1	Properties of <i>F</i> -Distribution	95
3.6.2	Continuous Random Variables	70		<i>Summary</i>	95
3.6.3	Probability Mass Function and Cumulative Distribution Function of a Discrete Random Variable	70		<i>Multiple Choice Questions</i>	96
3.6.4	Expected Value, Variance, and Standard Deviation of a Discrete Random Variable	71		<i>Exercises</i>	96
				<i>References</i>	98
3.7	Probability Density Function (PDF) and Cumulative Distribution Function (CDF) of a Continuous Random Variable	72	<b>4.</b>	<b>Sampling and Estimation</b>	<b>99</b>
3.8	Binomial Distribution	73	4.1	Introduction to Sampling	99
3.8.1	Probability Mass Function (PMF) of Binomial Distribution	74	4.1.1	When Jesus Christ became 2 <sup>nd</sup> Best in the World	101
3.8.2	Cumulative Distribution Function (CDF) of Binomial Distribution	74	4.2	Population Parameters and Sample Statistic	101
3.8.3	Mean and Variance of Binomial Distribution	75	4.3	Sampling	102
			4.4	Probabilistic Sampling	103
			4.4.1	Random Sampling	103
			4.4.2	Stratified Sampling	104
			4.4.3	Cluster Sampling	105
			4.4.4	Bootstrap Aggregating (Bagging)	105

<b>4.5</b>	Non-Probability Sampling	105	<b>6. Hypothesis Testing</b>	<b>133</b>	
4.5.1	Convenience Sampling	106	6.1	Introduction to Hypothesis Testing	134
4.5.2	Voluntary Sampling	106	6.1.1	Blackout Babies	134
<b>4.6</b>	Sampling Distribution	106	6.2	Setting Up a Hypothesis Test	135
<b>4.7</b>	Central Limit Theorem (CLT)	108	6.2.1	Description of Hypothesis	136
4.7.1	Central Limit Theorem for Proportions	109	6.2.2	Null and Alternative Hypothesis	136
<b>4.8</b>	Sample Size Estimation for Mean of the Population	110	6.2.3	Test Statistic	137
<b>4.9</b>	Estimation of Population Parameters	111	6.2.4	Decision Criteria – Significance Value	138
<b>4.10</b>	Method of Moments	113	6.3	One-Tailed and Two-tailed Test	138
<b>4.11</b>	Estimation of Parameters Using Method of Moments	114	6.4	Type I Error, Type II Error, and Power of The Hypothesis Test	141
<b>4.12</b>	Estimation of Parameters Using Maximum Likelihood Estimation	115	<b>6.5</b>	Hypothesis Testing for Population mean with Known Variance: $Z$ -Test	141
4.12.1	Estimation of Binomial Distribution Parameter	115	6.5.1	Power of Test and the Power Function	148
4.12.2	Estimation of Scale Parameter of Exponential Distribution	117	6.6	Hypothesis Testing for Population Proportion: $Z$ -Test for Proportion	149
4.12.3	MLE of Normal Distribution Parameters	118	6.7	Hypothesis Test for Population mean under Unknown Population Variance: $t$ -Test	151
<i>Summary</i>	119	6.8	Paired Sample $t$ -Test	154	
<i>Multiple Choice Questions</i>	119	6.9	Comparing Two Populations: Two-Sample $Z$ - and $t$ -Test	156	
<i>Exercises</i>	121	6.9.1	Difference in Two Population means when Population Standard Deviations are Known: Two-Sample $Z$ -Test	156	
<i>References</i>	122	6.9.2	Difference In Two Population Means When Population Standard Deviations are Unknown and Believed to be Equal: Two-sample $t$ -Test	158	
<b>5. Confidence Intervals</b>	<b>123</b>	6.9.3	Difference In Two Population Means When Population Standard Deviations are Unknown and not Equal: Two-sample $t$ -Test with Unequal Variance	159	
5.1	Introduction to Confidence Interval	123			
5.2	Confidence Interval for Population Mean	124			
5.3	Confidence Interval for Population Proportion	127			
5.4	Confidence Interval for Population Mean When Standard Deviation is Unknown	128			
5.5	Confidence Interval for Population Variance	129			
<i>Summary</i>	130				
<i>Multiple Choice Questions</i>	131				
<i>Exercises</i>	131				
<i>References</i>	132				

6.10 Hypothesis Test for Difference in Population Proportion under Large Samples: Two-Sample Z-Test for Proportions	161	8.2.1 Calculation of Pearson Product Moment Correlation Coefficient	209
6.11 Effect Size: Cohen's D	162	8.2.2 Spurious Correlation	213
6.12 Hypothesis Test for Equality of Population Variances	163	8.2.3 Hypothesis Test for Correlation Coefficient	213
6.13 Non-Parametric Tests: Chi-Square Tests	165	8.3 Spearman Rank Correlation	215
6.13.1 Chi-Square Goodness of Fit Tests	166	8.4 Point Bi-Serial Correlation	217
6.13.2 Choice of Number of Intervals in Chi-Square Goodness of Fit Test	172	8.5 The Phi-coefficient	218
6.13.3 Chi-Square Test of Independence	172	<i>Summary</i>	220
<i>Summary</i>	174	<i>Multiple Choice Questions</i>	220
<i>Multiple Choice Questions</i>	175	<i>Exercises</i>	220
<i>Exercises</i>	176	<i>References</i>	223
<i>Case Study: Central Parking Services Private Limited</i>	178		
<i>References</i>	186		
<b>7. Analysis of Variance</b>	<b>189</b>	<b>9. Simple Linear Regression</b>	<b>225</b>
7.1 Introduction to Analysis of Variance (ANOVA)	189	9.1 Introduction to Simple Linear Regression	226
7.2 Multiple <i>t</i> -Tests for Comparing Several Means	191	9.2 History of Regression—Francis Galton's Regression Model	228
7.3 One-way Analysis of Variance (ANOVA)	192	9.3 Simple Linear Regression Model Building	228
7.3.1 Setting Up an Analysis of Variance	192	9.4 Estimation of Parameters Using Ordinary Least Squares	234
7.3.2 Cochran's Theorem	194	9.5 Interpretation of Simple Linear Regression Coefficients	238
7.3.3 The <i>F</i> -Test	194	9.5.1 Interpretation of $\beta_0$ and $\beta_1$ in $Y = \beta_0 + \beta_1 X$	238
7.4 Two-Way Analysis of Variance (ANOVA)	199	9.5.2 Interpretation of $\beta_0$ and $\beta_1$ in $Y = \beta_0 + \beta_1 \ln(X)$	239
<i>Summary</i>	202	9.5.3 Interpretation of $\beta_0$ and $\beta_1$ in $\ln(Y) = \beta_0 + \beta_1 X$	239
<i>Multiple Choice Questions</i>	202	9.5.4 Interpretation of $\beta_0$ and $\beta_1$ In $\ln(Y) = \beta_0 + \beta_1 \ln(X)$	239
<i>Exercises</i>	203	9.6 Validation of the Simple Linear Regression Model	240
<i>References</i>	205	9.6.1 Coefficient of Determination ( <i>R</i> -Square or $R^2$ )	240
<b>8. Correlation Analysis</b>	<b>207</b>	9.6.2 Spurious Regression	242
8.1 Introduction to Correlation	207	9.6.3 Hypothesis Test for Regression Co-efficient ( <i>t</i> -Test)	243
8.2 Pearson Correlation Coefficient	208	9.6.4 Test For Overall Model: Analysis of Variance ( <i>F</i> -Test)	245
		9.6.5 Residual Analysis	246

9.7	Outlier Analysis	249	10.8	Validation of Multiple Regression Model	291
9.7.1	$Z$ -Score	250	10.9	Co-efficient of Multiple Determination ( $R$ -Square) and Adjusted $R$ -Square	291
9.7.2	Mahalanobis Distance	250	10.10	Statistical Significance of Individual Variables in MLR – $t$ -Test	292
9.7.3	Cook's Distance	250	10.11	Validation of Overall Regression Model: $F$ -Test	292
9.7.4	Leverage Value	250	10.12	Validation of Portions of a MLR Model – Partial $F$ -Test	293
9.7.5	DFFit and DFBeta	251	10.13	Residual Analysis in Multiple Linear Regression	294
9.8	Confidence Interval for Regression Coefficients $\beta_0$ and $\beta_1$	251	10.14	Multi-Collinearity and Variance Inflation Factor	295
9.9	Confidence Interval for the Expected Value of $Y$ for a Given $X$	252	10.14.1	Variance Inflation Factor (VIF)	295
9.10	Prediction Interval for the Value of $Y$ for a Given $X$	253	10.14.2	Remedies for Handling Multi-collinearity	296
	<i>Case Study: Package Pricing at the Die Another Day (DAD) Hospital Summary</i>	254	10.15	Auto-correlation	296
	<i>Multiple Choice Questions</i>	263	10.15.1	Durbin–Watson Test for Auto-correlation	297
	<i>Exercises</i>	264	10.16	Distance Measures and Outliers Diagnostics	297
	<i>References</i>	269	10.16.1	Mahalanobis Distance	298
<b>10.</b>	<b>Multiple Linear Regression</b>	<b>271</b>	10.16.2	Cook's Distance	298
10.1	Introduction	271	10.16.3	Leverage Value (or Hat Value)	299
10.2	Ordinary Least Squares Estimation for Multiple Linear Regression	272	10.16.4	DFFit and SDFFit	299
10.3	Multiple Linear Regression Model Building	275	10.16.5	DFBeta and SDFBeta	299
10.4	Part (Semi-Partial) Correlation and Regression Model Building	279	10.17	Variable Selection in Regression Model Building (Forward, Backward, and Stepwise Regression)	303
	10.4.1 Partial Correlation	279	10.17.1	Forward Selection	303
	10.4.2 Semi-Partial Correlation (or Part Correlation)	280	10.17.2	Backward Elimination Procedure	303
10.5	Interpretation of MLR Coefficients – Partial Regression Coefficient	282	10.17.3	Stepwise Regression	304
10.6	Standardized Regression Co-efficient	284	10.18	Avoiding Overfitting: Mallows's $C_p$ Transformations	305
10.7	Regression Models with Qualitative Variables	285	10.19	Tukey and Mosteller's Bulging Rule for Transformation	312
	10.7.1 Interpretation of Regression Coefficients of Categorical Variables	288		<i>Summary</i>	313
	10.7.2 Interaction Variables in Regression Models	289			

<i>Case Study: Pricing of Players in the Indian Premier League</i>	314	
<i>Multiple Choice Questions</i>	323	
<i>Exercises</i>	325	
<i>References</i>	335	
<b>11. Logistic Regression</b>	<b>337</b>	
11.1 Introduction – Classification Problems	337	
11.2 Introduction to Binary Logistic Regression	338	
11.3 Estimation of Parameters in Logistic Regression	340	
11.4 Interpretation of Logistic Regression Parameters	342	
11.5 Logistic Regression Model Diagnostics	343	
11.5.1 Omnibus Test (Likelihood Ratio Test)	344	
11.5.2 Wald's Test	346	
11.5.3 Hosmer–Lemeshow Test	346	
11.5.4 Pseudo $R^2$	347	
11.6 Classification Table, Sensitivity, and Specificity	347	
11.6.1 Accuracy Paradox	349	
11.6.2 Sensitivity, Specificity, and Precision	349	
11.6.3 Concordant and Discordant Pairs	350	
11.6.4 Receiver Operating Characteristics (ROC) Curve	352	
11.6.5 Area Under Roc Curve (AUC), Lorenz Curve, and Gini Coefficient	353	
11.7 Optimal Cut-Off Probability	354	
11.7.1 Classification Plot for Selection of Cut-Off Probability	354	
11.7.2 Youden's Index for Optimal Cut-Off Probability	356	
11.7.3 Cost-Based Cut-Off Probability	357	
11.8 Variable Selection in Logistic Regression	358	
11.8.1 Forward LR (Likelihood Ratio)	358	
11.8.2 Forward Selection Wald	358	
11.9 Application of Logistic Regression in Credit Rating	359	
11.9.1 Credit Score using Logistic Regression	362	
11.9.2 Youden's Index Calculation	363	
11.9.3 Classification Cut-Off Based on Penalty Cost	364	
11.10 Gain Chart and Lift Chart	365	
Summary	370	
Multiple Choice Questions	370	
Exercises	371	
<i>Case Study: HR Analytics at ScaleneWorks – Behavioural Modelling to Predict Renegue References</i>	381	
<i>References</i>	390	
<b>12. Decision Trees</b>	<b>391</b>	
12.1 Decision Trees: Introduction	391	
12.2 Chi-Square Automatic Interaction Detection (CHAID)	392	
12.2.1 CHAID Tree Development	393	
12.2.2 Bonferroni Correction	395	
12.2.3 Generating Business Rules using CHAID Tree	396	
12.3 Classification and Regression Tree	398	
12.3.1 Gini Impurity Index	399	
12.3.2 Entropy	401	
12.4 Cost-Based Splitting Criteria	401	
12.5 Ensemble Method	403	
12.6 Random Forest	404	
Summary	406	
Multiple Choice Questions	406	
Exercises	407	
<i>Case Study: Breaking Barriers: Micro-Mortgage Analytics References</i>	412	
<i>References</i>	426	

<b>13. Forecasting Techniques</b>	<b>427</b>	13.14 Auto-Regressive Integrated Moving Average (ARIMA) Process	463
13.1 Introduction to Forecasting	427	13.14.1 Dickey Fuller Test	464
13.2 Time-Series Data and Components of Time-Series Data	428	13.14.2 Augmented Dickey-Fuller Test	465
13.3 Forecasting Techniques and Forecasting Accuracy	430	13.14.3 Transforming Non-Stationary Process to Stationary Process using Differencing	465
13.3.1 Mean Absolute Error (MAE)	431	13.14.4 ARIMA( $p, d, q$ ) Model Building	466
13.3.2 Mean Absolute Percentage Error (MAPE)	431	13.14.5 Ljung–box Test for Auto-correlations	470
13.3.3 Mean Square Error (MSE)	431	13.15 Power of Forecasting Model: Theil's Coefficient	471
13.3.4 Root Mean Square Error (RMSE)	431	<i>Summary</i>	472
13.4 Moving Average Method	432	<i>Multiple Choice Questions</i>	472
13.5 Single Exponential Smoothing (ES)	435	<i>Exercises</i>	473
13.5.1 Optimal Smoothing Constant in a Single Exponential Smoothing (SES)	437	<i>Case Study: Larsen and Toubro – Spare Parts Forecasting</i>	477
13.6 Double Exponential Smoothing – Holt's Method	437	<i>References</i>	488
13.7 Triple Exponential Smoothing (Holt-Winter Model)	439	<b>14. Clustering</b>	<b>489</b>
13.7.1 Predicting Seasonality Index using Method of Averages	440	14.1 Introduction to Clustering	489
13.8 Croston's Forecasting Method for Intermittent Demand	441	14.2 Distance and Dissimilarity Measures used in Clustering	490
13.9 Regression Model for Forecasting	444	14.2.1 Euclidean Distance	490
13.9.1 Forecasting Time-series Data with Seasonal Variation	445	14.2.2 Standardized Euclidean Distance	493
13.10 Auto-Regressive (AR), Moving Average (MA) and ARMA Models	450	14.2.3 Manhattan Distance (City Block Distance)	493
13.11 Auto-Regressive (AR) Models	451	14.2.4 Minkowski Distance	493
13.11.1 AR Model Identification: ACF and PACF	452	14.2.5 Jaccard Similarity Coefficient (Jaccard Index)	493
13.12 Moving Average Process MA( $q$ )	457	14.2.6 Cosine Similarity	494
13.13 Auto-Regressive Moving Average (ARMA) Process	458	14.2.7 Gower's Similarity Coefficient	495
		14.3 Quality and Optimal Number of Clusters	496
		14.4 Clustering Algorithms	497
		14.4.1 Variable Selection	497
		14.4.2 Deciding Distance/Similarity Measures	497

14.4.3	Number of Clusters	498	15.13	Linear Integer Programming (ILP)	553
14.4.4	Cluster Validation	498	15.13.1	Branch and Bound Algorithm	554
14.5	K-Means Clustering	498	15.13.2	Branching Strategies In Branch and Bound Algorithm	558
14.6	Hierarchical Clustering	501	15.14	Multi-Criteria Decision-Making (MCDM) Problems	558
	<i>Summary</i>	504	15.14.1	Goal Programming	558
	<i>Multiple Choice Questions</i>	504			
	<i>Exercises</i>	505			
	<i>Case Study: Markdown Optimization for an Indian Apparel Retailer</i>	507			
	<i>References</i>	521			
<b>15.</b>	<b>Prescriptive Analytics</b>	<b>523</b>			
15.1	Introduction to Prescriptive Analytics	523			
15.2	Linear Programming	524			
15.3	Linear Programming (LP) Model Building	527	<b>16.</b>	<b>Stochastic Models</b>	<b>577</b>
15.4	Linear Programming Problem (LPP) Terminologies	530	16.1	Introduction Stochastic Process	578
15.5	Assumptions of Linear Programming	531	16.2	Poisson Process	578
15.6	Sensitivity Analysis in LPP	532	16.3	Compound Poisson Process	582
15.6.1	Change in the RHS of a Constraint	533	16.4	Markov Chains	583
15.6.2	Impact of Change in the Coefficient Values in Objective Function	534	16.4.1	One-Step Transition Probabilities of Markov Chain	584
15.6.3	100% Rule	534	16.4.2	Estimation of One-Step Transition Probabilities of Markov Chain	585
15.6.4	Addition of a new Constraint	535	16.4.3	Hypothesis Tests for Markov Chain: Anderson Goodman Test	585
15.6.5	Addition of a new Variable	535	16.4.4	Testing Time Homogeneity of Transition Matrices: Likelihood Ratio Test	589
15.7	Solving a Linear Programming Problem using Graphical Method	535	16.4.5	Using Markov Chains in Predictive Analytics	589
15.8	Range of Optimality	539	16.4.6	Stationary Distribution in a Markov Chain	590
15.9	Range of Shadow Price	540	16.4.7	Regular Matrix	592
15.10	Dual Linear Programming	540	16.5	Classification of States in a Markov Chain	594
15.10.1	Conversion of a primal Model to Dual Model	542	16.5.1	Accessible State	594
15.11	Primal–Dual Relationships	544	16.5.2	Communicating States	594
15.11.1	Weak Law of Duality	544	16.5.3	Recurrent and Transient States	594
15.11.2	Strong Law of Duality	545			
15.11.3	Complementary Slackness Theorem	545			
15.12	Multi-Period (Stage) Models	551			

16.5.4	First Passage Time and Mean Recurrence Time	595	17.8.2	Difference between Specification Limits and Control Limits	646
16.5.5	Periodic State	595	17.8.3	Importance of USL and LSL	646
16.5.6	Ergodic Markov Chain	596			
16.5.7	Limiting Probability	596			
16.6	Markov Chains with Absorbing States	596	17.9	Defects Per Million Opportunities (DPMO)	648
16.6.1	Canonical Form of the Transition Matrix of an Absorbing State Markov Chain	597	17.10	Yield	650
16.7	Expected Duration to Reach a State from other States	599	17.10.1	Rolled throughput Yield	651
16.8	Calculation of Retention Probability and Customer Lifetime Value using Markov Chains	601	17.11	Sigma Score (or Sigma Quality Level)	651
16.9	Markov Decision Process (MDP)	603	17.11.1	Conversion of Yield to Sigma Score under no Shift in the Process Mean	651
16.9.1	Policy Iteration Algorithm	605	17.11.2	Conversion of DPMO to Sigma Score under no Shift in Process Mean	652
16.9.2	Linear Programming Formulation for Finding Optimal Policy	608	17.11.3	Sigma Score under Process Shift	652
16.10	Value Iteration Algorithm	609	17.12	DMAIC Methodology	653
	<i>Summary</i>	612	17.12.1	Define Stage	654
	<i>Multiple Choice Questions</i>	612	17.12.2	Measure	654
	<i>Exercises</i>	613	17.12.3	Analyse	654
	<i>Case Study: Customer Analytics at Flipkart.com</i>	619	17.12.4	Improve	655
	<i>References</i>	631	17.12.5	Control	655
<b>17.</b>	<b>Six Sigma</b>	<b>633</b>	17.13	Six Sigma Project Selection For DMAIC Implementation	655
17.1	Introduction to Six Sigma	633	17.14	DMAIC Methodology – Case of Armoured Vehicle	655
17.2	What is Six Sigma?	636	17.15	Six Sigma Toolbox	661
17.3	Origins of Six Sigma	636	17.15.1	Cause-and-Effect Diagram	662
17.4	Three-Sigma versus Six-Sigma Process	638	17.15.2	SIPOC	662
17.5	Cost of Poor Quality	639	17.15.3	Five Whys	663
17.6	Sigma Score	640		<i>Summary</i>	664
17.7	Industrial Applications of Six Sigma	641		<i>Multiple Choice Questions</i>	664
17.8	Six Sigma Measures	643		<i>Exercises</i>	665
17.8.1	Process Capability and Process Capability Indices	643		<i>Case Study: Era of Quality at the Akshaya Patra Foundation</i>	666
				<i>References</i>	683
			<b>Appendix</b>		685
			<b>Bibliography</b>		705
			<b>Index</b>		707



# Introduction to Business Analytics

1

“This Exists, So that Exists This is not there, so that is not there  
This Ends, So that Ends This Arises, So that Arises.”

— The Buddha

## LEARNING OBJECTIVES

- LO 1-1** Learn foundations of analytics and how it is becoming a competitive strategy for many organizations.
- LO 1-2** Understand the importance of analytics in decision making and problem solving.
- LO 1-3** Understand how different organizations are using analytics to gain insights and add value.
- LO 1-4** Learn how organizations are using analytics to generate solutions and products.
- LO 1-5** Understand different types of analytical models such as descriptive analytics, predictive analytics, and prescriptive analytics.
- LO 1-6** Learn framework for analytics model development and deployment.
- LO 1-7** Understand frequently used tools and techniques in analytics and problems solved using such tools and techniques.

## BUSINESS ANALYTICS

Analytics has evolved from a simple number crunching exercise used for solving problems and assisting in decision making to a competitive strategy. In the beginning of the 21<sup>st</sup> century, analytics became one of the most important verticals within organizations due to its potential benefits including the ability to make better decisions and its impact on profitability of an organization. Today, several products and solutions are driven by analytics; Amazon Go, recommender systems, predictive keyboards used in smart phones and chatbot are few examples of solutions that are driven by analytics. It has become evident that analytics has become an important differentiator between high-performing and low-performing companies. Davenport and Patil (2012) claim that ‘data scientist’ will be the sexiest job of the 21<sup>st</sup> century.

IMPORTANT

*Analytics is not just about number crunching. It has evolved into a competitive strategy that drives innovation across several organizations.*

## 1.1 | INTRODUCTION TO BUSINESS ANALYTICS

In God we trust; all others must bring Data  
— Edwards Deming

The epigraph captures the importance of analytics and data-driven decision making in one sentence. During the early period of the 20<sup>th</sup> century, many companies were taking business decisions based on ‘opinions’ rather than decisions based on proper data analysis (which probably acted as a trigger for Deming’s quote). Opinion-based decision making can be very risky and often leads to incorrect decisions. One of the primary objectives of business analytics is to improve the quality of decision making using data analysis, which is the focus of this book.

Every organization across the world uses performance measures such as market share, profitability, sales growth, return on investment (ROI), customer satisfaction, and so on for quantifying, monitoring, benchmarking, and improving its performance. It is important for organizations to understand the association between key performance indicators (KPIs) and factors that have a significant impact on the KPIs for effective management. Knowledge of the relationship between KPIs and factors would provide the decision maker with appropriate actionable items. Analytics is a body of knowledge consisting of statistical, mathematical, and operations research techniques; artificial intelligence techniques such as machine learning and deep learning algorithms; data collection and storage; data management processes such as data extraction, transformation, and loading (ETL); and computing and big data technologies such as Hadoop, Spark, and Hive that create value by developing actionable items from data. Two primary macro-level objectives of analytics are problem solving and decision making. Analytics helps organizations to create value by solving problems effectively and assisting in decision making. Today, analytics is used as a competitive strategy by many organizations such as Amazon, Apple, General Electric, Google, Facebook and Procter and Gamble who use analytics to create products and solutions. Marshall (2016) and MacKenzie *et al.* (2013) reported that Amazon’s recommender systems resulted in a sales increase of 35%. Davenport and Harris (2007) and Hopkins *et al.* (2010) reported that there was a high correlation between use of analytics and business performance. They claimed that the majority of high performers (measured in terms of profit, shareholder return and revenue, etc.) strategically apply analytics in their daily operations, as compared to low performers.

Statistical and operations research techniques have been in use for several decades by many companies, but since 2000, companies that use analytics have increased exponentially. One reason for this increase in use of analytics is the *theory of bounded rationality* proposed by *Herbert Simon* (1972). According to Herbert Simon, the increasing complexity of business problems, the existence of several alternative solutions, and the limited time available for decision making demand a highly structured decision-making process using past data for the effective management of organizations. Decision making has become difficult due to reasons such as uncertainty, incomplete information about alternatives, lack of knowledge about cause and effect relationships between parameters of importance, and time available for decision making. For example, fraudulent transactions are a major problem for e-commerce companies; one such fraud is customers returning fake items in place of genuine items that they purchased (for example, buying branded Ray-Ban sunglasses and returning a fake sunglasses). Once the customer returns an item, the e-commerce company, such as Amazon, processes the refund of money within 3 to 5 business days (source: Amazon website). Given such a time constraint, e-commerce companies have to

identify fraudulent transactions in real time or within a very short duration (check whether the returned item is fake or genuine and start the refund process). But it is easier said than done since an expert is required to differentiate a fake product from a genuine one. What makes this problem even more difficult is that the number of stock keeping units (SKUs) sold by e-commerce companies runs into several millions and the number of transactions can run into several millions in a day making it a highly complex problem to deal with. Valentina Palladino (2013) reported that Amazon sold 426 items per second prior to December 2013 Christmas. The scale of operations of 21<sup>st</sup> century companies is huge and makes it difficult to manage the business without analytics. In 2015, Flipkart sold over 30 million products from more than 50,000 sellers through their platform. The number of visits to their portal was more than 10 million daily and the number of shipments exceeded 8 million per month (Bhansali *et al.*, 2016). A few of the problems that e-commerce companies such as Amazon and Flipkart try to address are as follows:

1. Forecasting demand for products directly sold by the company; excess inventory and shortage can impact both the top line and the bottom line.
2. Cancellation of orders placed by customers before their delivery. Ability to predict cancellations and intervention can save cost incurred on unnecessary logistics.
3. Fraudulent transactions resulting in financial loss to the company.
4. Predicting delivery time since it is an important service level agreement from the customer perspective.
5. Predicting what a customer is likely to buy in future to create recommender systems.

Given the scale of operations of modern companies, it is almost impossible to manage them effectively without analytics. Although decisions are occasionally made using the HiPPO algorithm (“highest paid person’s opinion” algorithm), especially in a group decision-making scenario, there is a significant change in the form of “data-driven decision making” among several companies. Many companies use analytics as a competitive strategy and many more are likely to follow. A typical data-driven decision-making process uses the following steps (Figure 1.1):

1. Identify the problem or opportunity for value creation.
2. Identify sources of data (primary as well secondary data sources).
3. Pre-process the data for issues such as missing and incorrect data. Generate derived variables and transform the data if necessary. Prepare the data for analytics model building.
4. Divide the data sets into subsets training and validation data sets.
5. Build analytical models and identify the best model(s) using model performance in validation data.
6. Implement Solution/Decision/Develop Product.

Analytics is used to solve a wide range of problems starting with simple process improvement such as reducing procurement cycle time to complex decision-making problems such as farm advisory systems that involve accurate weather prediction, forecasting commodity price etc, so that farmers can be advised about crop selection, crop rotation, etc. Figure 1.2 shows the pyramid of analytics applications, at the bottom of the pyramid analytics is used for process improvement and at the top it is used for decision making and as a competitive strategy.

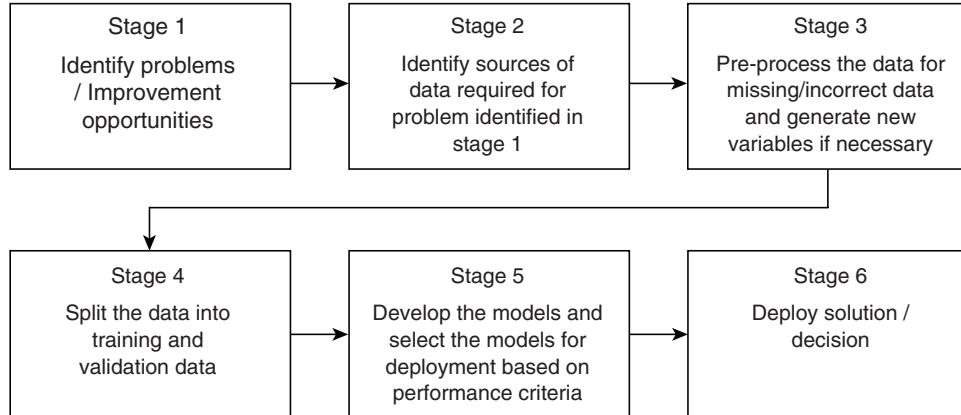


FIGURE 1.1 Business analytics – Data-driven decision-making flow diagram.

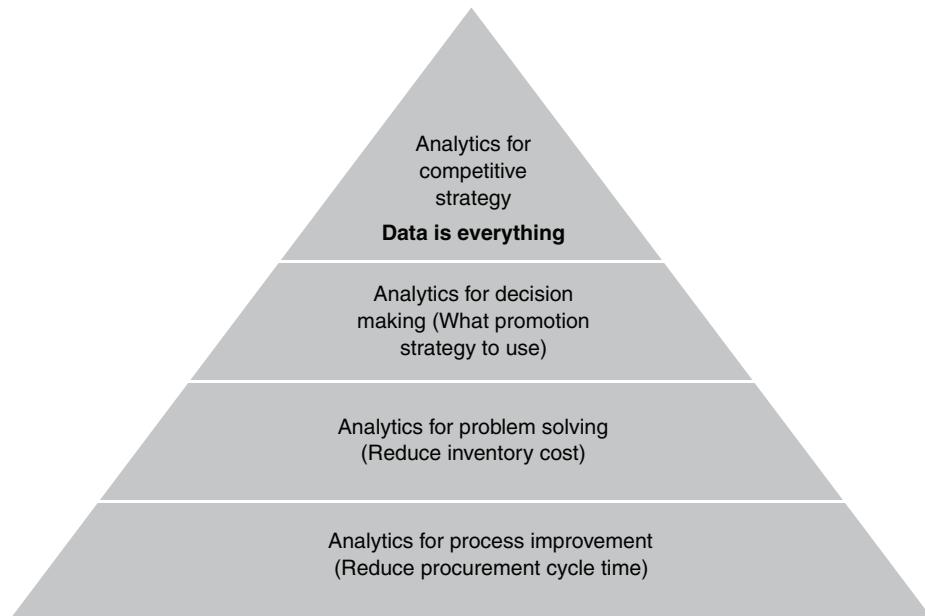


FIGURE 1.2 Pyramid of analytics.

Ransbotham and Kiron (2017) reported that they observed an increasing trend in companies using analytics to drive innovation and several companies reported competitive advantage from use of analytics.

## 1.2 | WHY ANALYTICS

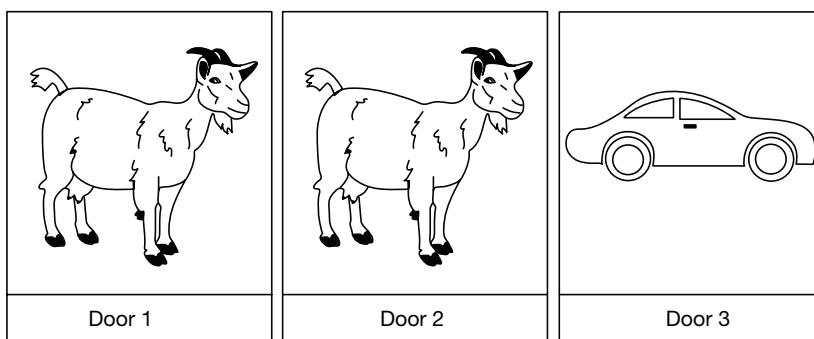
According to the *theory of firm* (Coase, 1937 and Fame, 1980) as proposed by several economists, firms exist to minimize the transaction cost. Transactions take place when goods or services are transferred to

customers from the supplier. The cost of decision making is an important element of transaction cost. Michalos (1970) groups the costs of decision making into three categories:

1. Cost of reaching a decision with the help of a decision maker or procedure; this is also known as production cost, that is, cost of producing a decision.
2. Cost of actions based on decisions produced; also known as implementation cost.
3. Failure costs that account for failure of an organization's efforts on production and implementation.

The profit earned by the firm would depend on how well they manage to minimize the transaction costs. Profit maximization or transaction cost minimization would mean making right decisions about the market, product/service, processes, supply chain, and so on. For example, consider a firm that would like to sell product such as a ready made shirt. The firm has to take several decisions such as fabric, colour, size, fit, price, promotion, and so on. Each of these attributes has several options. The real problem starts with decision-making ability of firms, especially the techniques and processes used in decision making; unfortunately human beings are inherently not good at decision making. A great example for human's inability to take decisions is the famous *Monty Hall problem* (Savant, 1990) in which the contestants of a game show are shown three doors (Figure 1.3). Behind one of the doors is an expensive item (such as a car or gold); while there are inexpensive items behind the remaining two doors (such as a goat). The contestant is asked to choose one of the doors. Assume that the contestant chooses door 1; the game host would then open one of the remaining two doors. Assume that the game host opens door 2, which has a goat behind it. Now the contestant is given a chance to change his initial choice (from door 1 to door 3). The problem is whether or not the contestant should change his/her initial choice. Note that the contestant is given an option to switch door irrespective of the item behind his/her original choice of door. The problem is based on a famous television show "Let's make a deal" hosted by Monty Hall in 1960s and 1970s (Selvin, 1975).

In this problem, the contestant — the decision maker — has two choices: he/she can either change his/her initial choice or stick with his/her initial choice. When Marilyn vos Savant, a columnist at the *Parade Magazine*, posted that the contestant should change the initial choice (Savant, 1990), 92% of the general public and 65% of the university graduates (many of them with PhDs) who responded to her column were against her answer.<sup>1</sup> Although Marilyn vos Savant provided a simple decision tree



**FIGURE 1.3** Monty Hall problem.

<sup>1</sup> Source: [http://en.wikipedia.org/wiki/Monty\\_Hall\\_problem](http://en.wikipedia.org/wiki/Monty_Hall_problem)

argument to prove that the probability of winning increases to  $2/3$  when the contestant changes his/her initial choice, many scholars did not accept her argument that changing the initial option is the right decision. Table 1.1 shows why changing the initial option increases the probability of winning. The expensive item can be behind any one of the three doors as shown in Table 1.1 (rows 2–4). Assume that the contestant has chosen door 1 initially, columns 4 and 5 (last row) give the probability of winning the car if contestant stays with door 1 (column 4) and the door 1 is changed (column 5), respectively.

The above argument can be extended to any number of doors without loss of generality. In the case of Monty Hall problem, the number of alternatives available to the player is just two. Even when the number of options is only 2, many find it difficult to comprehend that changing the initial choice will increase the probability of winning. In many real-life decision-making scenarios, the number of options available to the decision maker can be several millions or billions. The travelling salesman problem (TSP) is one such decision-making problem which many companies encounter in their business. In a TSP, given a list of cities and the distances between each pair of cities, the objective is to find the shortest possible route a salesperson should take to visit each city exactly once and return to the origin city. Many organizations that need to deliver products to many locations on regular basis encounter TSP. For example, in 2015 the Akshaya Patra Foundation (TAPF), which provides mid-day meals to approximately 1.4 million under privileged school children across India, faced this decision-making challenge (Srujana *et al.*, 2015). In 2015, through their Vasanthapura kitchen in Bangalore, approximately 84000 school children from 650 schools in South Bangalore were provided mid-day meals. Providing high quality food at an affordable price is one of the challenges faced by Akshaya Patra. The Vasanthapura kitchen used 35 vehicles to distribute the cooked food (Srujana *et al.*, 2015). To minimize the cost of distribution, they need to solve a complex vehicle routing problem (VRP). To simplify this problem, assume that they divide the number of schools equally among the vehicles; each vehicle would then have to deliver food to approximately 20 schools (few vehicles are kept as standby). For each vehicle, we need to find the best route. This problem can be formulated as a TSP with a solution space of 20 factorial ( $20! = 2.4329 \times 10^{18}$ ). If a computer can evaluate one million routes per second, it would take more than 77146 years to evaluate all possible routes! For Akshaya Patra, every rupee saved would enable them to add more children to their mid-day meal programme. Given that the human brain lacks the ability to take the right decision in the Monty Hall problem that has just two alternatives, a problem with 20 factorial alternatives is certainly beyond the human brain's processing ability. With approximately 270 employees, Akshaya Patra's kitchen in Vasanthapura falls under the category of small- and medium-size enterprises (SMEs). Despite being an SME, it has to handle a complex analytics problem that many believe is relevant

**TABLE 1.1** Monty Hall problem final probability of win when the player changes the initial choice

Item Behind Door 1	Item Behind Door 2	Item Behind Door 3	Result if Stayed with Door #1	Result if the Door is Changed
Car	Goat	Goat	Car	Goat
Goat	Car	Goat	Goat	Car
Goat	Goat	Car	Goat	Car
Probability of Winning		1/3		2/3

only for big organizations. One of the misconceptions about analytics and big data technologies is that it is appropriate only for large organizations; however, the truth is that any organization, small or big, can benefit from the use of analytics. TSP is a problem that is encountered by several e-commerce companies for delivery of items placed by the customers and logistics service providers.

In today's world, data-driven decision making through business analytics is just not an option, but an essential capability that every organization should acquire irrespective of its size. As the competition increases, organizations cannot afford to shield inefficiencies. Analytics provides the capability for the organizations to be efficient and effective. Based on a survey of 3000 executives, Hopkins *et al.* (2010) claimed that there is a striking correlation between an organization's analytics sophistication and its competitive performance. The biggest obstacle to adopting analytics is the lack of knowledge about the tools and techniques that are required.

### 1.3 | BUSINESS ANALYTICS: THE SCIENCE OF DATA-DRIVEN DECISION MAKING

“Go down deep enough into anything and you will find Mathematics”

— Dean Schlicter

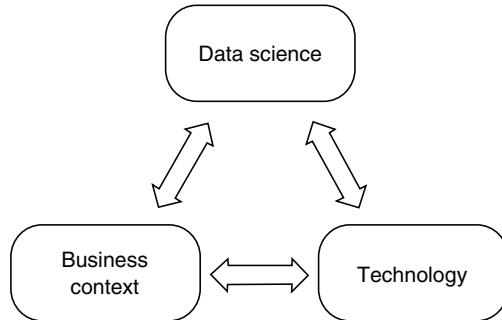
Business analytics is a set of statistical and operations research techniques, artificial intelligence, information technology and management strategies used for framing a business problem, collecting data, and analysing the data to create value to organizations.

Increasing complexities associated with businesses in the form of scale of operations and competition demand deeper understanding of the market and customers to serve better and succeed in the market. One of the main reasons for analytics is the scale of operations. If Walmart was a country, its GDP would be 28<sup>th</sup> in the world, its revenue in 2014 was 485.7 billion US dollars (Snyder, 2015). Merchandizing, shelf space allocation, promotions, brand monitoring, managing talent at the scale of operations of Walmart, Target, and Amazon requires solving complex problems in real time. The human mind lacks the ability to choose the right decisions due to the complexity of the problems that the organizations are facing and the limited time available for decision making (Simon, 1972).

In the 1980s, the culture of data collection was poor. Many organizations did not collect data or the data collected was not in a form that could be easily used for deriving insights. Even in 2017, many companies collect data manually which may result in data quality issues. Organizations found decision making difficult due to the lack of data that could be made available quickly. The introduction of enterprise resource planning (ERP) systems in many organizations partially solved the problem of non-availability of data that can be called upon whenever needed. However, the data sitting in the ERP systems needed to be analysed for problem solving and decision making; the original ERP systems were not designed to build analytics models. Platforms such as SAP HANA and Microsoft Azure try to fill this gap.

Business Analytics can be broken into 3 components:

1. Business Context
2. Technology
3. Data Science



**FIGURE 1.4** Components of business analytics.

Interaction between the three components is shown in Figure 1.4.

### 1.3.1 | Business Context

Business analytics projects start with the business context and ability of the organization to ask the right questions. Consider Target's Pregnancy prediction (Duhigg, 2012), which is a great example for organizations' ability to ask the right questions. Target is one of the largest retail chains in the world and in 2015, the revenue of Target Corporation was approximately US \$ 73 billion. According to Duhigg (2012), Target developed a model to assign a pregnancy score to each female customer among their shoppers which could be further used for target marketing. However, what is so special about this prediction, and why pregnant women? This is where the knowledge of business context plays an important role. The following business knowledge manifests the importance of pregnancy prediction from a retailer's perspective:

1. Pregnant women are likely to be price-insensitive, so they become the Holy Grail for retailers such as Target. Expectant women are usually willing to spend more for their comfort as well as the babies'.
2. US Department of Agriculture reported that the expenses on children in 2015 ranged between US Dollars (USD) 9,600 and USD 19,700 (Lino *et al.*, 2017). According to National Vital Statistics report (2017), close to 4 million children were born in 2015, that is, the market size of baby-related products was at least USD 38 billion (Martin *et al.*, 2017). The market size was probably similar during the early 2000s when Target developed the pregnancy prediction model.
3. For many customers shopping is a habit, and most do not respond to promotions since shopping is a routine for them. The shopping behaviour changes during special events such as marriage and pregnancy and it becomes easy to target them during these special events.

The 'pregnancy prediction' is based on the insights about price-insensitive customers and the market size of baby products. In analytics, knowledge of business context is important for the ability to ask the right questions to start the analytics project.

Another good example of business context driving analytics is the 'did you forget feature' used by the Indian online grocery store bigbasket.com (Abraham *et al.*, 2016). Many customers have the tendency to forget items they intended to buy. Fernandes *et al.* (2013) reported that on average, customers forget

30% of the items they intend to buy. Forgetfulness can have significant cost impact for the online grocery stores. The customers may buy the forgotten items from a nearby store where they live, but since she/he is already in the store she/he may buy more items resulting in reduction in basket size in the future for online grocery stores such as bigbasket.com. Alternatively, the customer may place another order for forgotten items, but this time, the size of the basket is likely to be small and results in unnecessary logistics cost. Thus, the ability to predict the items that a customer may have forgotten to order can have a significant impact on the profits of online grocers such as bigbasket.com.

Another problem that online grocery customers face while ordering the items is the time taken to place an order. Unlike customers of Amazon or Flipkart, online grocery customers order several items each time; the number of items in an order may cross 100. Searching for all the items that a customer would like to order is a time-consuming exercise, especially when they order using smart phones. Thus, bigbasket created a ‘smart basket’ which is a basket consisting of items that a customer is likely to buy (recommended basket) reducing the time required to place the order (Abraham *et al.*, 2016).

The above two examples (Target’s pregnancy test and ‘did you forget’ and smart basket feature at bigbasket.com) manifest the importance of business context in business analytics, that is, the ability to ask the right questions is an important success criteria for analytics projects.

### 1.3.2 | Technology

To find out whether a customer is pregnant or to find out whether a customer has forgotten to place an order for an item, we need data. In both the cases, the point of sale data has to be captured consisting of past purchases made by the customer. Information Technology (IT) is used for data capture, data storage, data preparation, data analysis, and data share. Today most data are unstructured data; data that is not in the form of a matrix (rows and columns) is called unstructured data. Images, texts, voice, video, click stream are few examples of unstructured data. To analyse data, one may need to use software such as R, Python, SAS, SPSS, Tableau, etc. Technology is also required to deploy the solution; for example, in the case of Target, technology can be used to personalize coupons that can be sent to individual customers. An important output of analytics is automation of actionable items derived from analytical models; automation of actionable items is usually achieved using IT.

### 1.3.3 | Data Science

Data Science is the most important component of analytics, it consists of statistical and operations research techniques, machine learning and deep learning algorithms. Given a problem, the objective of the data science component of analytics is to identify the most appropriate statistical model/machine learning algorithm that can be used. For example, Target’s pregnancy prediction is a *classification problem* in which customers (or entities) are classified into different groups. In the case of pregnancy test, the classes are:

1. Pregnant
2. Not pregnant

There are several techniques available for solving classification problems such as logistic regression, classification trees, random forest, adaptive boosting, neural networks, and so on. The objective of the data science component is to identify the technique that is best based on a measure of accuracy. Usually, several models are developed for solving the problem using different techniques and a few models may be chosen for deployment of the solution.

Business analytics can be grouped into three types: **descriptive analytics**, **predictive analytics**, and **prescriptive analytics**. In the following sections, we shall discuss the three types of analytics in detail.

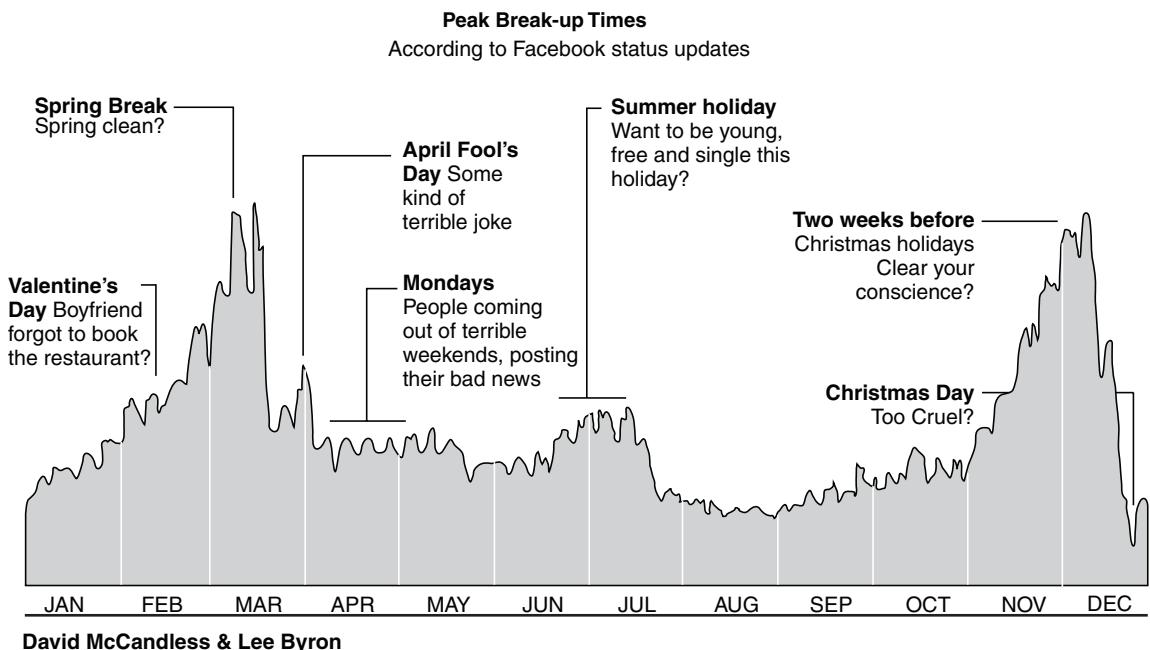
## 1.4 | DESCRIPTIVE ANALYTICS

“If the statistics are boring, then you’ve got the wrong numbers”.

— Edward R. Tufte

Descriptive analytics is the simplest form of analytics that mainly uses simple descriptive statistics, data visualization techniques, and business related queries to understand past data. One of the primary objectives of descriptive analytics is innovative ways of data summarization. Descriptive analytics is used for understanding the trends in past data which can be useful for generating insights. Figure 1.5 shows visualization of relationship break-ups reported in Facebook.

It is clear from Figure 1.5 that spike in breakups occurred during spring break and in December before Christmas. There could be many reasons for increase in breakups during December (we hope it is



David McCandless & Lee Byron  
InformationisBeautiful.net / LeeByron.com

Source: Facebook Lexicon 2008

FIGURE 1.5 Peak breakup times according to Facebook status update. Source: David McCandless and Lee Byron.

not a New Year resolution that they would like to change the partner). Many believe that since December is a holiday season, couples get a lot of time to talk to each other, probably that is where the problem starts. However, descriptive analytics is not about why a pattern exists, but about what the pattern means for a business. The fact that there is a significant increase in breakups during December we can deduce the following insights (or possibilities):

1. There will be more traffic to online dating sites during December/January.
2. There will be greater demand for relationship counsellors and lawyers.
3. There will be greater demand for housing and the housing prices are likely to increase in December/January.
4. There will be greater demand for household items.
5. People would like to forget the past, so they might change the brand of beer they drink.

Descriptive analytics using visualization identifies trends in the data and connects the dots to gain insights about associated businesses. In addition to visualization, descriptive analytics uses descriptive statistics and queries to gain insights from the data. The following are a few examples of insights obtained using descriptive analytics reported in literature:

1. Most shoppers turn towards the right side when they enter a retail store (Underhill, 2009, pages 77–79). Retailers keep products with higher profit on the right side of the store since most people turn right.
2. Married men who kiss their wife before going to work live longer, earn more and get into less number of accidents as compared to those who do not (Foer, 2006).
3. Correlated with Facebook relationship breakups, divorces spike in January. According to Caroline Kent (2015), January 3 is nicknamed ‘divorce day’.
4. Men are more reluctant to use coupons as compared to women (Hu and Jasper, 2004). While sending coupons, retailers should target female shoppers as they are more likely to use coupons.

Trends obtained through descriptive analytics can be used to derive actionable items. For example, when Hurricane Charley struck the U.S. in 2004, Linda M. Dillman, Walmart's Chief Information Officer, wanted to understand the purchasing behaviour of their customers (Hays, 2004). Using data mining techniques, Walmart found that the demand for strawberry pop-tarts went up over 7 times during the hurricane compared to their normal sales rate; the pre-hurricane top-selling item was found to be beer. These insights were used by Walmart when the next hurricane — Hurricane Frances — hit the U.S. in August–September 2004; most of the items predicted by Walmart sold quickly. Although the high pre-hurricane demand for beer can be intuitively predicted, the demand for strawberry pop-tarts was a complete surprise.

Data visualization to understand hidden facts/trends has been in use for several centuries. Dr. John Snow's cholera spot map is an interesting application of data visualization. Cholera claimed millions of lives across the world during the 19th century. Medical practitioners did not have a clear understanding of the causes of the disease (Cameron and Jones, 1983). Between 1845 and 1856, over 700 articles were

published in London on the causes of cholera and how the epidemic could be prevented (Snow, 2002). However, none of them offered any real solution. The breakthrough in cholera epidemiology was made by Dr. John Snow based on the data of cholera outbreak in central London in 1854. Between 31 August and 10 September 1854, over 500 people died of cholera in London. John Snow marked the locations of the homes of those who had died during the epidemic and prepared a spot map (Figure 1.6)<sup>2</sup>. The spot map revealed that the highest number of deaths occurred in the Golden Square area (Snow, 1999). The most striking difference between this area and the other districts of London was the source of water (Brody *et al.*, 2000); thus, Snow established that water contamination was the main source of cholera.

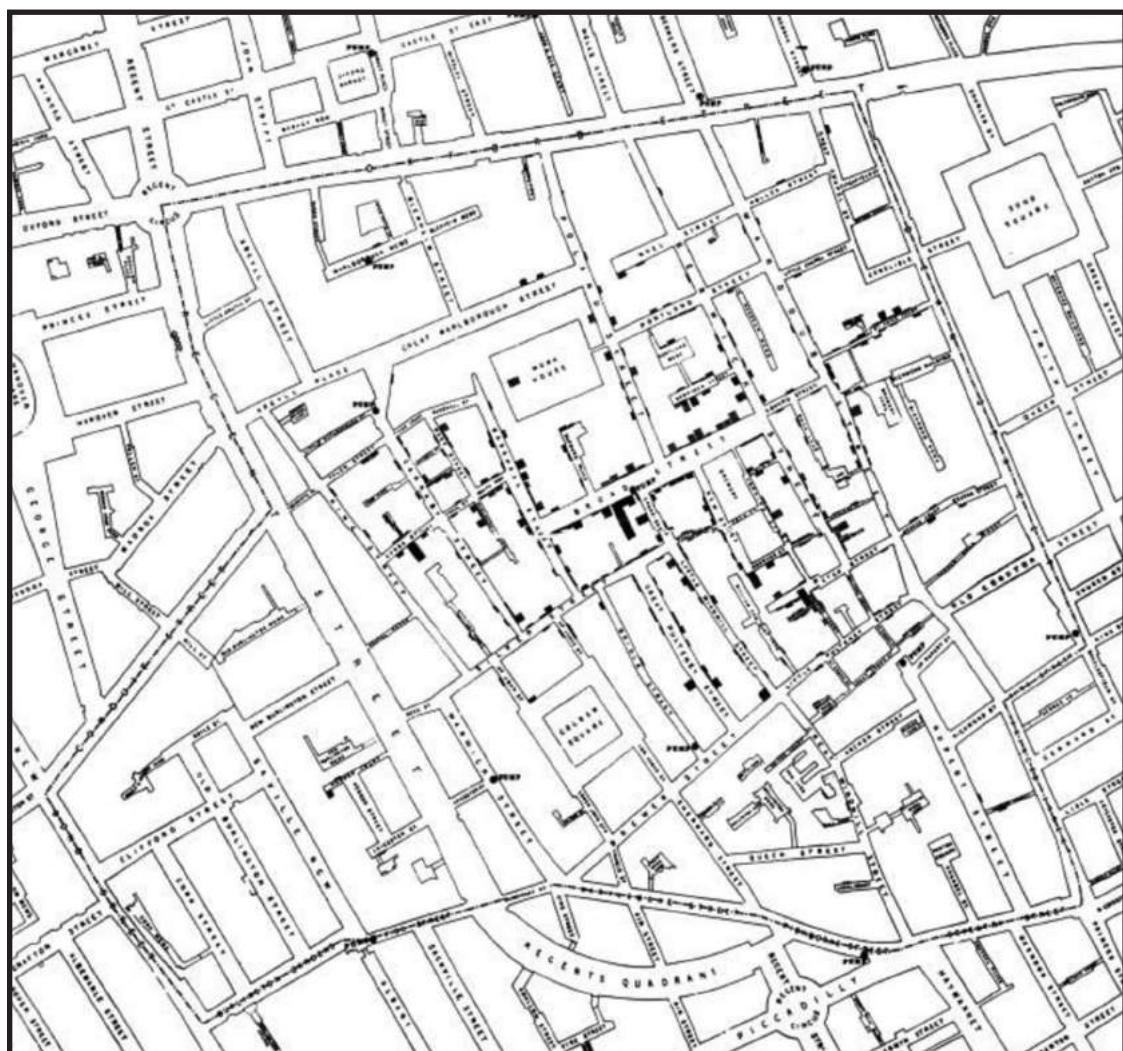


FIGURE 1.6 John Snow's spot map of cholera outbreak in London, 1854.

<sup>2</sup> Source: [http://en.wikipedia.org/wiki/John\\_Snow\\_\(physician\)](http://en.wikipedia.org/wiki/John_Snow_(physician))

Edward Tufte (2001), in his book titled *The Visual Display of Quantitative Information*, demonstrated how innovative visuals can be used to effectively communicate data. Google search keywords are used to predict demand for different apparel styles, jewellery, footwear, and so on to understand demand trends for many products. These trends help retailers take better decisions regarding procurement and inventory planning. Dashboards are created using innovative visuals form the core of business intelligence and are an important element of analytics. Tableau and Qlik Sense are popular visualization tools that are used by several organizations to create dashboards to monitor several key performance indicators relevant for the organization in real time. Indian companies such as Gramener<sup>3</sup> have used innovative data visualization tools to communicate hidden facts in the data. Descriptive analytics could be the initial stage of creating analytics capability.

Simple analysis of data can lead to business practices that result in financial rewards. For instance, companies such as RadioShack and Best Buy found a high correlation between the success of individual stores and the number of female employees in the sales team (Underhill, 2009). Underhill (2009) also reported that the conversion rate (percentage of people who purchased something) in consumer durable shops was higher among female shoppers than among male shoppers. Many organizations across the globe have to deal with fraudulent transactions. Sometimes, a simple query can lead to fraud detection. In 2014, China Eastern Airline found that a man had booked a first class ticket more than 300 times in a year and cancelled it before its expiry for full refund so that he could eat free food at the airport's VIP lounge (David K Li, 2014). In India, insurance frauds accounted for 2500–3500 crore in 2010 (Anon, 2013). It is always a good practice to start analytics projects with descriptive analytics.

## 1.5 | PREDICTIVE ANALYTICS

If you torture the data long enough, it will confess.

— Ronald Coase

In the analytics capability maturity model (ACMM), predictive analytics comes after descriptive analytics and is the most important analytics capability. It aims to predict the probability of occurrence of a future event such as forecasting demand for products/services, customer churn, employee attrition, loan defaults, fraudulent transactions, insurance claim, and stock market fluctuations. While descriptive analytics is used for finding what has happened in the past, predictive analytics is used for predicting what is likely to happen in the future. The ability to predict a future event such as an economic slowdown, a sudden surge or decline in a commodity's price, which customer is likely to churn, what will be the total claim from auto insurance customer, how long a patient is likely to stay in the hospital, and so on will help organizations plan their future course of action. Anecdotal evidence suggests that predictive analytics is the most frequently used type of analytics across several industries. The reason for this is that almost every organization would like to forecast the demand for the products that they sell, prices of the materials used by them, and so on. Irrespective of the type of business, organizations would like to forecast the demand for their products or services and understand the causes of demand fluctuations. The use of predictive analytics can reveal relationships that were previously unknown and are not intuitive.

<sup>3</sup> Source: <https://gramener.com/>

The most popular example of the application of predictive analytics is Target's pregnancy prediction model discussed earlier in the chapter. In 2002, Target hired statistician Andrew Pole; one of his assignments was to predict whether a customer is pregnant (Duhigg, 2012). At the outset, the question posed by the marketing department to Pole may look bizarre, but it made great business sense. Any marketer would like to identify the price-insensitive customers among the shoppers, and who can beat soon-to-be parents? A list of interesting applications of predictive analytics is presented in Table 1.2.

The examples shown in Table 1.2 represent a tiny fraction of the predictive analytics applications used in the industry. Companies such as Procter & Gamble use analytics as a competitive strategy — every critical management decision is made using analytics (Davenport, 2013). If one were to search for the reasons behind highly successful companies, one would usually find analytics being deployed as the competitive strategy. Google — without which many people think the world would end — uses Markov chains for page ranking (Hayes, 2013). Google also developed accurate prediction models that could predict events such as the outcome of political elections, the launch date of a product, or action(s) taken by competitors (Coles *et al.*, 2007). Davenport and Harris (2007) reported how companies such as Amazon, Capital One, Harrah's, and the Boston Red Sox have dominated their business by using analytics. The application of analytics is not restricted to big corporates only; many sports clubs have successfully used analytics to manage their clubs. The most famous application of analytics in sports is by Oakland Athletics, which used analytics to put together a team with the limited resources available

**TABLE 1.2** List of predictive analytics applications

Organization	Predictive Analytics Model
Polyphonic HMI	Predicts whether a song will be a hit using machine learning algorithms. Their product 'Hit Song Science' uses mathematical and statistical techniques to predict the success of a song on a scale of 1 to 10 (Anon, 2003).
Okcupid	Predicts which online dating message is likely to get a response from the opposite sex (Siegel, 2013).
Amazon.com	Uses predictive analytics to recommend products to their customers. It is reported that 35% of Amazon's sales is achieved through their recommender system (Siegel, 2013, MacKinzie <i>et al.</i> , 2013).
Hewlett Packard (HP)	Developed a flight risk score for its employees to predict who is likely to leave the company (Siegel, 2013).
University of Maryland	Claimed that dreams can predict whether one's spouse will cheat (Whitelocks, 2013).
Flight Caster	Predicts flight delays 6 hours before the airline's alerts.
Netflix	Predicts which movie their customer is likely to watch next (Greene, 2006). 75% of what customer watch at Netflix is from product recommendations (MacKinzie <i>et al.</i> , 2013).
Capital One Bank	Predicts the most profitable customer (Davenport, 2007).
Google	Predicted the spread of H1N1 flu using the query terms (Carneiro and Mylonakis, 2010).
Farecast	Developed a model to predict airfare, whether it is likely to increase or decrease, and the amount of increase/decrease. <sup>4</sup>

<sup>4</sup> Source: <http://www.crunchbase.com/company/farecast>

for purchasing players (Lewis, 2003). Oakland Athletics had the third lowest payroll among the major league baseball teams in 2002. The manager of Oakland Athletics, Billy Beane, used statistical techniques to identify player qualities that made an impact on the match outcome and to also identify relatively cheaper skill. Oakland Athletics revised their team management strategy and with a payroll of USD 41 million, they were able to compete successfully in the league. In 2002, they won 20 games in a row.

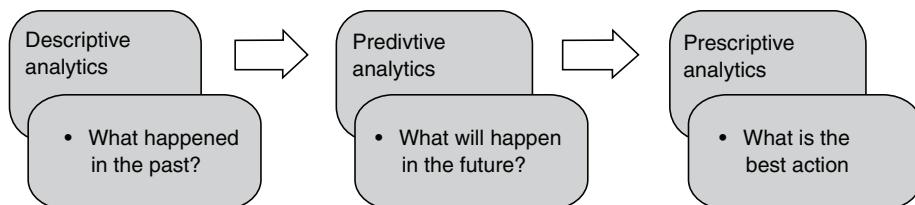
1.6 | PRESCRIPTIVE ANALYTICS

Every decision has a consequence.

— Damon Darrel

Prescriptive analytics is the highest level of analytics capability which is used for choosing optimal actions once an organization gains insights through descriptive and predictive analytics. In many cases, prescriptive analytics is solved as a separate optimization problem. Prescriptive analytics assists users in finding the optimal solution to a problem or in making the right choice/decision among several alternatives. Operations Research (OR) techniques form the core of prescriptive analytics. Apart from operations research techniques, machine learning algorithms, metaheuristics, and advanced statistical models are used in prescriptive analytics. Note that actionable items can be derived directly after descriptive and predictive analytics model development; however, they may not be the optimal action. For example, in a Business to Business (B to B) sales, the proportion of sales conversions to sales leads could be very low. The sales conversion period could be very long, as high as 6 months to one year. Predictive analytics such as logistics regression can be used for predicting the propensity to buy a product and actionable items (such as which customer to target) can be derived directly based on predicted probability to buy or using lift chart. However, the values of the sale are likely to be different, as are the profits earned from different customers. Thus, targeting customers purely based on probability to buy may not result in an optimal solution. Use of techniques such as binary integer programming will result in optimal targeting of customers that maximize total expected profit. That is, while actionable items can be derived from descriptive and predictive analytics, use of prescriptive analytics ensures optimal actions (choices or alternatives). The link between different analytics capability is shown in Figure 1.7.

Ever since their introduction during World War II, OR models have been used in every sector of every industry. The list of prescriptive analytics applications is long and several companies across the world have benefitted from the use of prescriptive analytics tools. Coca-Cola Enterprises (CCE) is the largest distributor of Coca-Cola products. In 2005, CCE distributed 2 billion physical cases



**FIGURE 1.7** Link between different analytics capabilities.

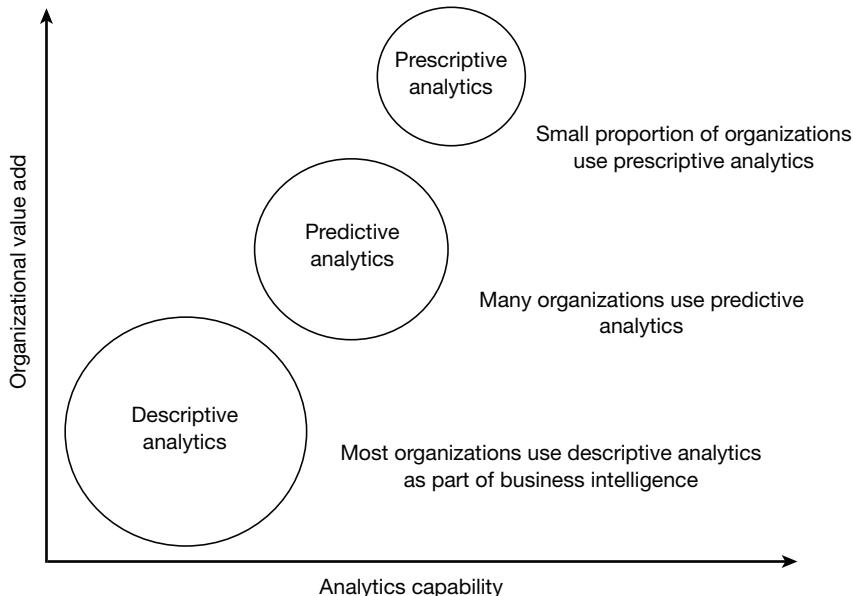


FIGURE 1.8 Analytics capability versus organizational value add.

containing 42 billion bottles and cans of Coca-Cola (Kant *et al.*, 2008). CCE developed an OR model that would meet several objectives such as improved customer satisfaction and optimal asset utilization for their distribution network of Coca-Cola products from 430 distribution centres to 2.4 million retail outlets. The optimization model resulted in cost savings of USD 54 million and improved customer satisfaction. The Akshaya Patra Midday Meal Routing and Transportation Algorithm (AMRUTA) was developed to solve the vehicle routing problem (discussed in Section 1.1); this was implemented at Akshaya Patra's Vasanthpura campus, resulting in savings of USD 75000 per annum (Mahadevan *et al.*, 2013). A major challenge for any e-commerce company is to improve the conversion of visits to transactions and order sizes. Hewlett Packard (HP) established HPDirect.com in 2005 to build online sales. HP Global Analytics developed predictive and prescriptive analytics techniques to improve sales. The analytical solutions helped HP to increase conversion rates and order sizes (Rohit *et al.*, 2013).

Inventory management is one of the problems that is most frequently addressed using prescriptive analytics. Samsung implemented a set of methodologies under the title '*Short Life and Low Inventory in Manufacturing*' (SLIM) to manage all the manufacturing and supply chain problems. Between 1996 and 1999, Samsung implemented SLIM in all its manufacturing facilities, resulting in a reduction in the manufacturing cycle time of random access memory devices from more than 80 days to less than 30 days. SLIM enabled Samsung to capture additional markets worth USD 1 billion (Leachman *et al.*, 2002). Product mix, marketing mix, travelling salesman problem, vehicle routing, facility location, manpower planning, capital budgeting, transportation and capacity management are a few frequently addressed prescriptive analytics problems. Figure 1.8 shows analytics capability and organizational value add; prescriptive analytics provides maximum value add to the organization since the benefits are realized during every period.

## 1.7 | DESCRIPTIVE, PREDICTIVE, AND PRESCRIPTIVE ANALYTICS TECHNIQUES

The most frequently used predictive analytics techniques are regression, logistic regression, classification trees, forecasting, K-nearest neighbours, Markov chains, random forest, boosting, and neural networks. The frequently used tools in prescriptive analytics are linear programming, integer programming, multi-criteria decision-making models such as goal programming and analytic hierarchy process, combinatorial optimization, non-linear programming, and meta-heuristics. In Table 1.3, we provide a brief description of some of these tools and the problems that are solved using these tools. We have highlighted a few tools that are frequently used by analytics companies.

**TABLE 1.3** Predictive and prescriptive analytics techniques

Analytics Techniques	Applications
Regression	Regression is the most frequently used predictive analytics tool. It is a supervised learning algorithm. In management and social sciences, almost all hypotheses are validated using regression models. In business, irrespective of the sector, the decision maker would like to know how the key performance indicators (KPIs) of the business are related to macro-economic parameters and other internal process parameters. Regression is an excellent tool for establishing the existence of an association relationship between a response variable (KPI) and other explanatory variables. Unfortunately, regression is one of the most highly misused techniques in analytics.
Logistic and Multinomial Regression	Logistic and multinomial logistic regression techniques are used to find the probability of occurrence of an event. Logistic regression is a supervised learning algorithm. Logistic regression is used for solving classification and discrete choice problems. Classification problems are common in many businesses. For example, banks and financial institutions would like to classify their customers into several risk categories. Companies would like to predict which customer is highly likely to churn in the next quarter. Marketers would like to know which brand a customer is likely to buy and whether promotions can make a customer change his/her brand loyalty. Credit scoring and fraud detection are other popular applications of logistic regression.
Decision Trees	Decision trees or classification trees are usually used for solving classification problems. There are several types of classification tree models. Chi-Squared Automatic Interaction Detection (CHAID) and Classification Trees (CART) are frequently used for solving classification problems. Although the decision trees are usually used for solving classification problems (in which the outcome variable is discrete), they can also be used when the outcome variable is continuous.
Markov Chains	Olle Haggstrom (2007) wrote an article stating that problem solving is often a matter of cooking up an appropriate Markov chain. One of the initial applications of Markov chains was implemented by the American retail giant Sears. They used a Markov Decision Process to decide the optimal mailing policy for their catalogues (Howard, 2002). Today, Markov chains are one of the key analytics tools in marketing, finance, operations, and supply chain management.
Random Forest	Random forest is one of the popular machine learning algorithms that uses ensemble approach to solve the problem by generating a large number of models.
Linear Programming	Since its origins during World War II, linear programming is one of the most frequently used techniques in prescriptive analytics. Problems such as resource allocation, product mix, cutting-stock problem, revenue management, and logistics optimisation are frequently solved using linear programming.
Integer Programming	Many optimization problems in real life may have variables that can take only integer values. When one or more variables in the problem can take only an integer solution, the model is called an integer programming model. Capital budgeting, scheduling, and set covering are a few problems that are solved using integer programming.

(Continued)

**TABLE 1.3** (Continued)

Analytics Techniques	Applications
Multi-Criteria Decision-Making Model	In many cases, the decision makers may have more than one objective (or KPIs). For example, a company may like to increase the profit as well as the market share. It is possible that the various objectives identified by the organization may conflict with one another. In such cases, techniques such as Analytic Hierarchy Process and Goal Programming are used to arrive at the optimal decisions.
Combinatorial Optimisation	Combinatorial optimization involves choosing the optimal solution from a large number of finite solutions. The travelling salesman problem (TSP), the vehicle routing problem (VRP), and the minimum spanning tree problem (MST) belongs to this category. Many industry problems are analogical to TSP, VRP, and MST.
Non-Linear Programming (NLP)	Large classes of problems faced by the industry have non-linear objective functions and/or non-linear constraints. Many engineering design optimization problems belongs to this category. NLP are also difficult set of problems to solve due to limitations of the existing algorithms. NLP is an integral part of several machine learning algorithms such as neural networks. The loss function which is used for finding weights for input variables is a non-linear function.
Six Sigma	Six Sigma and its problem-solving methodology DMAIC (Define, Measure, Analyse, Improve, and Control) are frequently used in process improvement problems.
Social Media Analytics Tools	Social media analytics is a collection of tools and techniques used for analysing unstructured data such as texts, videos, photos, and so on. With the exponential growth in the use of social media by the general public, tools designed for analysing unstructured data will be frequently used by organizations.

## 1.8 | BIG DATA ANALYTICS

The world is one big data problem.

— Andrew McAfee

Big data is a class of problems that challenge existing IT and computing technology and existing algorithms. Traditionally, big data is defined as a big volume of data (in excess of 1 terabyte) generated at high velocity with high variety and veracity. That is, big data is identified using 4 Vs, namely, volume, velocity, variety, and veracity which are defined as follows:

1. Volume is the size of the data that an organization holds. Typically, this can run into several petabytes ( $10^{15}$  bytes) or exabytes ( $10^{16}$  bytes). Organizations such as telecom and banking collect and store a large quantity of customer data. Data collected using satellite and other machine generated data such as data generated by health and usage monitoring systems fitted in aircrafts, weather and rain monitoring systems can run into several exabytes since the data is captured minute by minute.
2. Velocity is the rate at which the data is generated. For example, AT&T customers generated more than 82 petabytes of data traffic on a daily basis (Anon, 2016).
3. Variety refers to the different types of data collected. In the case of telecom, the different data types are voice calls, messages in different languages, video calls, use of Apps, etc.

4. Veracity of the data refers to the quality issues. Especially in social media there could be biased or incorrect data, which can result in wrong inferences.

Big data is mostly misused terminology since a large proportion of analytics projects are not big data projects, in the sense that they do not challenge the existing computing technology and algorithms. Many companies see big data as a technology problem. Although it is true that technology is a constraint while addressing big data problems, a true big data problem challenges the algorithms that we have today, that is the algorithms are inadequate to solve these problems within a reasonable time. However, many problems in predictive and prescriptive analytics belongs to the big data category. When the problem associated with the data can challenge the existing computing technology due to its volume, velocity, and variety, then we have a big data problem. For example, Google processes 24 petabytes of data every day (Mayer-Schonberger and Cukier, 2013). Google was the first to exploit big data for targeted advertising using clickstream data. Google also predicted the spread of H1N1 flu based on the search terms entered by Google users (Ginsberg *et al.*, 2009). According to Richard Kellet, Director of Marketing at SAS, we (human) created 500 times more data in the last 10 years than what we had done prior to that, since the beginning of humanity (Scott, 2013). Every B787 flight creates half a terabyte of machine-generated data. For large banks, the automatic teller machine (ATM) transactions themselves will run into several billion transactions per month. Telecom call data, social media data, banking and financial transactions, machine-generated data, and healthcare data are a few examples of the sources of big data. Alternatively, any problem that can challenge the existing computing power, IT systems, and/or algorithms constitutes a big data problem. Big data problems need innovative ideas in order to handle the 4 Vs for deriving insights. The volume of the data generated is increasing every day, and most of this data is user generated, mainly from social media platforms, or machine generated. With the increase in Internet penetration and autonomous data capturing, the velocity of data is also increasing at a faster rate. It is estimated that 2.5 exabyte of data is created every day; this figure is likely to double in the near future (Mayer-Schonberger and Cukier, 2013). As the velocity of the data increases, traditional models such as regression and classification techniques may become unstable and invalid for analysis. The variety of data is another challenge within big data. Even in the case of text mining, users may use different languages within the same sentence, especially in a country such as India, which is home to hundreds of languages. Natural language processing (NLP), which is an essential component of big data, is challenging when multiple languages are used in the text data.

Innovative parallel processing capabilities such as Apache Hadoop (that comes with the ability to process large-scale data sets in multiple clusters using the Hadoop Distributed File System) and MapReduce (which enables parallel processing of large data sets) are used by organizations to handle big data. Big data technologies are still in a nascent stage and are far from maturity. However, these big data technologies do provide better computing power compared to existing technologies.

## 1.9 | WEB AND SOCIAL MEDIA ANALYTICS

Social media and mobile devices such as smart phones are becoming an important source of data for all organizations, small and big. Social media is also an important marketing channel for marketers since it

helps to create a buzz or electronic word-of-mouth (WoM) effectively. Stelzner (2013) claimed that 86% of the marketers indicated that social media is important for their business. Stelzner (2013) identified the following questions as the most relevant for any marketers when dealing with social media engagement (also valid for mobile devices):

1. What is the most effective social media tactic?
2. What are the best ways to engage the customers with social media?
3. How to calculate the return on investment on social media engagement?
4. What are the best social media management tools?
5. How do we create a social media strategy for the organization?

From the literature on social media and interviews conducted by IIMB (Dinesh Kumar *et al*, 2014) with industry experts, it is evident that social media is important for marketing products and services. However, the effectiveness of the social media marketing is still an emerging subject. Suhruta *et al.* (2013) claimed that there is a relationship between social media engagement and the box-office collection of movies based on the data obtained from the Bollywood movie '1920 evil returns'. Social media has several advantages over conventional media as discussed below. Social media is measurable in terms of impressions, visits, views, clicks, comments, shares, likes, followers, fans, subscribers, etc. Impact of conventional media cannot be measured, for example, views of a hoarding or newspaper ad cannot be measured.

Social media is less expensive than conventional media and has the potential to reach a wider audience. Social media can create viral impact in a short duration and can reach a larger number of people. A key challenge in social media strategy will be assessing the return on investment. Return on Investment (ROI) should be calculated by the formula

$$\text{ROI} = \frac{\text{(Gain from Social Media Marketing} - \text{Cost of Social Media Marketing})}{\text{Cost of Social Media Marketing}}$$

However, it is difficult to quantify the actual gain from social media marketing. Hence, several variations are used to calculate ROI as given below (Hemann and Burbary, 2013, pages 276-284):

1. **Return on Engagement (ROE):** This measures the impact of social media marketing on users' engagement on the premise that higher engagement leads to higher awareness and thus greater likeliness to make a purchase decision (Hemann and Burbary, 2013). ROE calculation for some of the social media platforms is provided below:
  - Facebook – (Number of likes, comments, and shares on a post)/(Total number of Facebook page likes)
  - Twitter – (Number of replies, re-tweets)/(Number of followers)
  - YouTube – (Number of comments, ratings, and likes)/(Number of video views) OR (Number of comments, ratings, and likes)/(Number of subscribers)
2. **Return on Influence:** This tries to measure how social media activity changes the behaviour of users.
3. **Anecdotes:** This measures verbal sharing of sales activity or intent of purchase on the social media platforms.

4. **Correlation:** This measures the relationship between any social media engagement activity and actual sales.
5. **Multivariate Testing:** This measures the relationship between multiple social media engagement activities and actual sales and enables providing the right kind of offers and promotions to different users.
6. **Linking and Tagging:** This approach provides links on the social media to the buyers to make their purchase and thus it is possible to relate sales and social media engagement. Another way is to embed 'Cookies' (a piece of software), which track consumers' online activity, thus providing the connect between social media engagement and actual sales. However, this approach is more effective when the sales are conducted online.
7. **Social Commerce Approach:** In this method, sales are directly conducted through social media; for example, a store front is set up on Facebook page.
8. **Share of Conversation:** (Volume of conversation for a particular brand)/(Volume of conversation for entire industry)
9. **Sentiment Analysis:** Tracks overall brand perception by crawling through all the data available on the net. Kumar and Mirchandani (2012) have proposed new measures such as customer influence effect (CIE), stickiness index, and customer influence value (CIV).

## 1.10 | MACHINE LEARNING ALGORITHMS

Machine learning algorithms are part of artificial intelligence (AI) that imitates the human learning process. Humans learn through multiple experiences to perform a task. Similarly, machine learning algorithms usually develop multiple models and each model is equivalent to an experience. For example, consider someone trying to learn tennis. Getting the service right requires much practice especially, to learn to serve an ace (serve such that the opponent player is unable to reach the ball). To master the service in tennis (especially ace), a player probably has to practice several thousand times; each practice session is a learning. AI is still a developing field and nowhere near the human cognitive process. In machine learning algorithms, we develop several models which can run into several hundreds and each model is treated as a learning opportunity. Mitchell (2006) defined machine learning as follows:

"Machine learns with respect to a particular task  $T$ , performance metric  $P$  following experience  $E$ , if the system reliably improves its performance  $P$  at task  $T$ , following experience  $E$ ".

Let the task  $T$  be a classification problem. To be more specific, consider customer's propensity to buy a product. The performance  $P$  can be measured through several metrics such as overall accuracy, sensitivity, specificity, and area under the receive operating characteristic curve (AUC). The experience  $E$  is analogous to different classifiers generated in machine learning algorithms such as random forest (in random forest several trees are generated and each tree is used for classification of new case). Carbonell *et al.* (1983) list the following three dimensions of machine learning algorithms:

1. Learning strategies used by the system.
2. Knowledge or skill acquired by the system.
3. Application domain for which the knowledge is obtained.

Carbonell *et al.* (1983) classifies learning into two groups: knowledge acquisition and skill refinement. They give an example of knowledge acquisition as learning concepts in physics whereas skill refinement is similar to learning to play the piano or ride a bicycle. Machine learning algorithms imitate both knowledge acquisition as well as skill refinement process. Machine learning algorithms are classified into the following four categories:

1. **Supervised Learning Algorithms:** When the training data set has both predictors (input) and outcome (output) variables, we use supervised learning algorithms. That is, the learning is supervised by the fact that predictors ( $X$ ) and the outcome ( $Y$ ) are available for the model to use. Techniques such as regression, logistic regression, decision tree learning, random forest, and so on are supervised learning algorithms.
2. **Unsupervised Learning Algorithms:** When the training data has only predictor (input) variables ( $X$ ), but not the outcome variable ( $Y$ ), then we use unsupervised learning algorithms. Techniques such as K-means clustering and Hierarchical clustering are examples of unsupervised learning algorithms.
3. **Reinforcement Learning Algorithms:** In many cases, the input variable  $X$  and the output variable  $Y$  are uncertain (predictive keyboards and/or spell check). The algorithms are also used in sequential decision-making scenarios; techniques such as dynamic programming and Markov decision process are examples of reinforcement learning algorithms.
4. **Evolutionary Learning Algorithms:** These are algorithms that imitate human/animal learning process. They are most frequently used to solve prescriptive analytics problems. Techniques such as genetic algorithms and ant colony optimization belongs to this category.

In this book, we will be discussing techniques from supervised, unsupervised, and reinforcement learning algorithms.

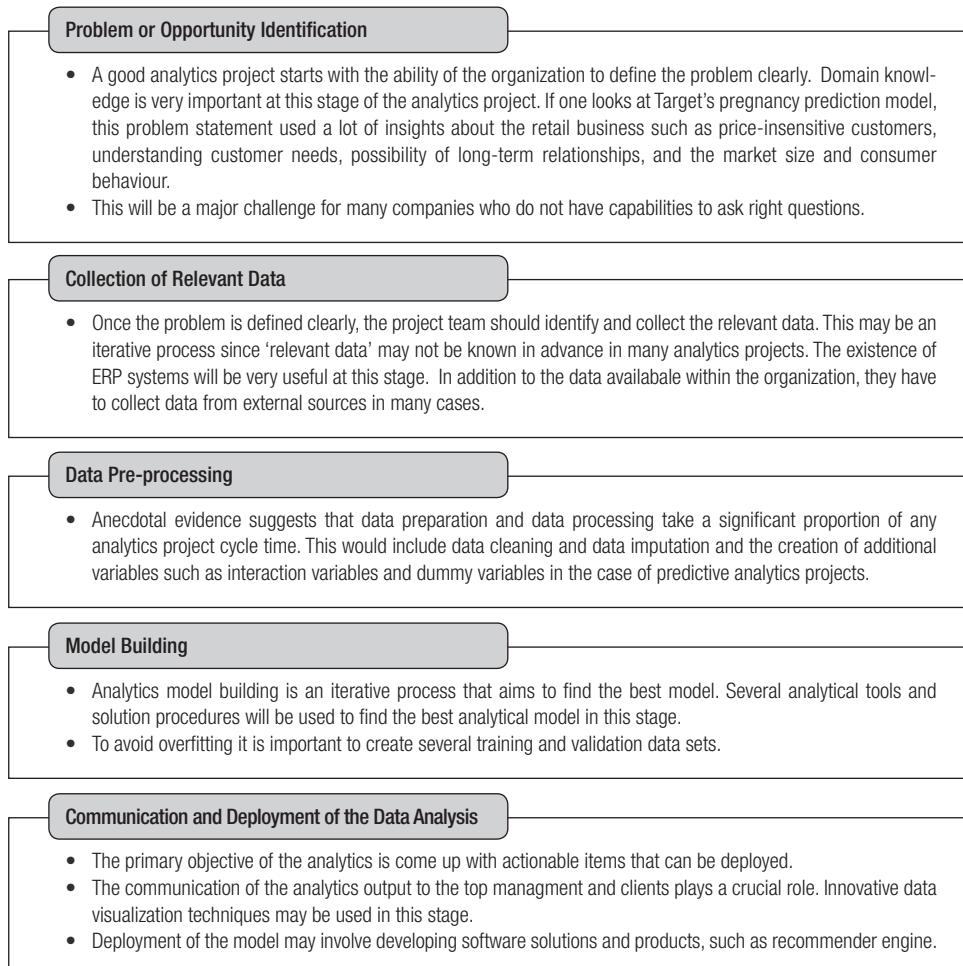
## 1.11 | FRAMEWORK FOR DATA-DRIVEN DECISION MAKING

The framework for data-driven decision making and problem solving can be divided into five integrated stages: *problem and opportunity identification; collection of relevant data; data pre-processing; analytics model building; and model deployment*. The various activities carried out during these different stages are described in Figure 1.9. The success of analytics projects will depend on how innovatively the data is used by the organization as compared to the mechanical use of analytical tools. Although there are several routine analytics projects such as customer segmentation, clustering, forecasting, and so on, highly successful companies blend innovation with analytics.

## 1.12 | ANALYTICS CAPABILITY BUILDING

Although many companies have successful analytics verticals within their organization, many are still in the process of creating one. In this section, we will discuss the pillars of building a centre of analytics excellence.

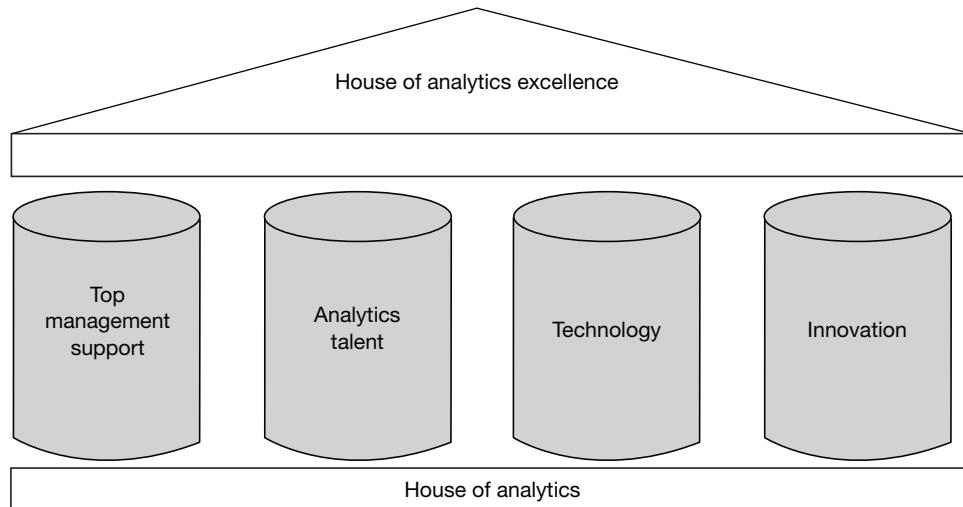
1. **Top Management Support:** Like many other initiatives, creating a data-driven decision-making process requires a change in the organizational culture, and without the support of the top-level



**FIGURE 1.9** Framework for data-driven decision making.

management, it may be difficult to create a strong analytics culture within an organization. Data-driven decision making may not result in immediate benefits, especially when the process is new to an organization.

2. **Analytics Talent:** The second important factor in creating a successful analytics vertical is the talent. It is important that organizations identify the right talent and nurture them within the organization to avoid attrition. The organization should have the ability to differentiate the true analytics talent from the mediocre analytics professionals. Davenport and Patil (2012) listed 10 ideas for finding the right data scientists such as recruiting from top universities, using social media such as LinkedIn, looking for evidence, and so on.
3. **Information Technology (IT):** IT plays a crucial role in implementing analytics. Data capturing, data storage, data transfer, data analysis through analytical models, and finally, communication of the model output cannot be achieved without proper data architecture supported by other IT infrastructure. In addition to data handling capabilities, software tools such as R, Python, SAS,



**FIGURE 1.10** House of analytics excellence.

SPSS, STATA, Tableau, and so on are an essential part of IT support. A large number of organizations, including multinational companies, prefer open-source software such as R instead of proprietary software.<sup>5</sup> Organizations expect real-time decisions; thus, in-memory computing is becoming popular among analytics companies.

4. **Innovation:** The fourth pillar of analytics excellence is innovation. Without innovation, the analytics function may not achieve its full potential. All these pillars need to be integrated with the domain knowledge of the business; otherwise, the analytics may end up solving non-value-adding problems. The house of analytics is shown in Figure 1.10.

### 1.13 | ROADMAP FOR ANALYTICS CAPABILITY BUILDING

Many organizations, whether small or big, have a large number of low-hanging fruits that can be targeted with simple analytics tools such as descriptive statistics, data visualization, pivot tables, correlation analysis, basic quality tools, lean and Six Sigma. Data summarization tools such as pivot tables of Microsoft Excel can be used for targeting several small improvement opportunities. Lean and Six Sigma concepts are usually a good way to start analytics practices in an organization if they do not have strong analytics expertise and would like to initiate analytics practice within their organization. Companies who are planning to start analytics divisions to support decision making may use the framework shown in Figure 1.11. Sample industry-wide applications of analytics are captured in Table 1.4.

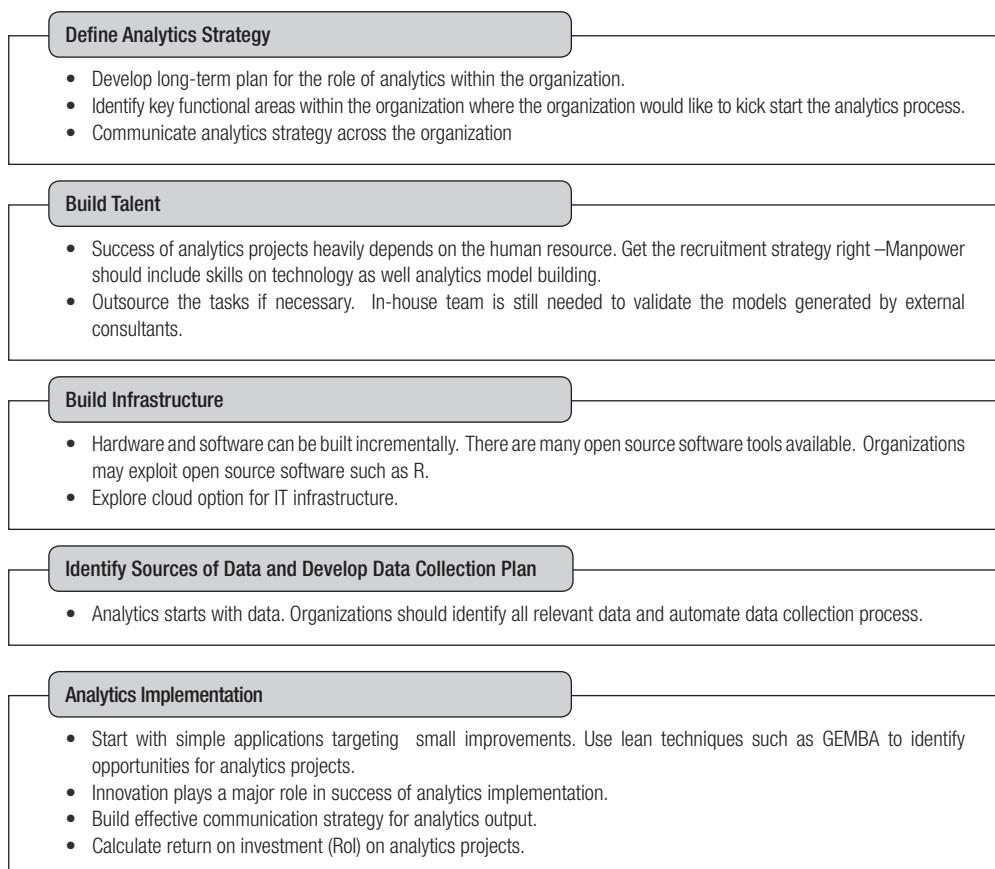
In addition to the primary data, the companies need to use data from secondary sources such as social media and other data sources such as Centre for Monitoring Indian Economy (CMIE), Capitaline, Bloomberg, Thomson Reuters, A C Nielson, Indiastat.com, etc.

<sup>5</sup> Many universities across the world teach analytics using R due to its capability, and not simply because it is open-source software.

## 1.14 | CHALLENGES IN DATA-DRIVEN DECISION MAKING AND FUTURE

Analytics requires a cultural change in an organization and managing this change will be the most significant challenge for many companies. This is true for any new initiative and is not specific to analytics. However, unlike many other initiatives, developing analytics skills can be a major hurdle, if these skills are not already present in the organization. Employees who are not skilled in analytical tools would need to be trained. Unlike other training programmes, analytics training can be long and expensive. Sirkin *et al.* (2005) identified duration, performance integrity, commitment, and effort as important factors that would determine the outcome of the change initiative.

If analytics talent is not available internally, the company should establish a system to identify the right talent. Building the right talent pool and retaining the talent would be a key challenge. Another big challenge would be the investment — the IT infrastructure required for advanced analytical techniques can be expensive. However, small and medium enterprises can achieve significant improvements by using simple tools such as MS Excel and open source software such as R and Python.



**FIGURE 1.11** Roadmap for analytics capability building.

**TABLE 1.4** Examples of industry-wise analytical problems and data resources

Industry Sector	Sample Analytical Problems	Data Sources
Manufacturing	<p><b>Supply Chain Analytics:</b> Inventory management, procurement, vendor selection, distribution management</p> <p><b>Quality and Process Improvement:</b> Product quality, manufacturing quality, process improvement</p> <p><b>Revenue and Cost Management:</b> Revenue maximization and cost minimization.</p> <p><b>Warranty Analytics:</b> Manage end customer warranty and after sales support.</p>	<ul style="list-style-type: none"> <li>▪ Procurement, sales and production data.</li> <li>▪ Warranty and after sales service data.</li> <li>▪ Commodity price data</li> <li>▪ Manufacturing data.</li> <li>▪ Macroeconomic data.</li> </ul>
Retail	<p><b>Assortment Planning:</b> Category and SKU (stock keeping unit) management that will maximize the revenue and improve loyalty.</p> <p><b>Promotion Planning:</b> Decide promotion strategy such as temporary price cuts, markdowns, bundling, etc.</p> <p><b>Demand Forecasting:</b> Forecast demand at SKU level for managing supply chain.</p> <p><b>Market Basket Analysis:</b> Association among SKUs in customer purchase.</p> <p><b>Customer Segmentation:</b> Identify the customer segmentation for target marketing.</p>	<ul style="list-style-type: none"> <li>▪ Price data.</li> <li>▪ Demand data at SKU and at category level.</li> <li>▪ SKU level sales data with and without promotions.</li> <li>▪ Planogram</li> <li>▪ Customer demographics data.</li> <li>▪ Point of Sales (PoS) data.</li> <li>▪ Loyalty program data.</li> </ul>
Healthcare	<p><b>Clinical Care:</b> Data related to clinical care and treatment required for improving quality of care.</p> <p><b>Hospitality related data:</b> Data related to issues such as registration process, housekeeping, nursing, utility, diagnostics, etc.</p>	<ul style="list-style-type: none"> <li>▪ All patient care related data</li> <li>▪ Hospitality related data</li> <li>▪ Patient feedback data</li> </ul>
Service	<p><b>Demand Forecasting:</b> Forecast demand for the service.</p> <p><b>NPS Optimization:</b> Net Promoters Score is an important measure of organizational performance.</p> <p><b>Service Quality Analysis:</b> Analyse quality for benchmarking and improvement.</p> <p><b>Customer Segmentation:</b> Used for target marketing.</p> <p><b>Promotion:</b> Promotion and its impact.</p>	<ul style="list-style-type: none"> <li>▪ Transactional and feedback data</li> <li>▪ Pricing and demand data</li> <li>▪ Promotional data</li> </ul>
Banking and Finance	<p><b>Service Demand Analysis:</b> Demand for different services.</p> <p><b>Customer Transaction Analysis:</b> Used for many different analytics and decision-making insights.</p> <p><b>Credit Scoring:</b> Important for managing different portfolios.</p>	<ul style="list-style-type: none"> <li>▪ Customer transactional data</li> <li>▪ Loan originating data</li> <li>▪ Credit scoring data</li> </ul>
IT and ITES (IT enables services)	<p><b>Demand for Analytics Services:</b> Identify demand for analytics products and services,</p> <p><b>Software Development Cycle Time:</b> Cost and time reduction.</p>	<ul style="list-style-type: none"> <li>▪ Customer interaction and market research data</li> <li>▪ Internal product development data</li> </ul>

Analytics will become an integral part of organizations and majority of the decisions will be made using data in the future. Innovation will be the key success factor for analytics deployment.

## 1.15 | ORGANIZATION OF THE BOOK

The focus of the book is on data science. The book starts with basic concepts in statistics such as descriptive statistics, axioms of probability, concept of random variables, discrete and continuous probability distributions, hypothesis testing, and correlations. After the introduction to the basic concepts in probability and statistics, advanced concepts such as regression, logistic regression, decision trees, forecasting, and clustering are introduced with practical examples and case studies. Final few chapters are dedicated to prescriptive analytics, stochastic models, and Six Sigma.

We will be discussing several examples and case studies throughout the book for better understanding of the concepts discussed. Exercise questions from Chapter 9 onwards require deeper understanding of the concepts. Case studies provided in this chapter are distributed through Harvard Business Publishing case portal and are used by many institutions across the world.

The data sets used in the book can be downloaded from the following website:

<https://www.wileyindia.com/business-analytics-the-science-of-data-driven-decision-making.html>

The readers may go through the predictive analytics course offered by the author on edX platform. The course is part of the Indian Institute of Management Bangalore's massive open online course (MOOC). The course videos are available at the following link:

<https://www.edx.org/course/predictive-analytics-iimbx-qm901x>

## REFERENCES

1. Anon (2003), “Major Music Labels Use Artificial Intelligence to help determine Hitability of Music”, Music Industry News Network, 25 February 2003.
2. Anon (2013), “Fraud in Motor and Health Insurance Global Perspective: Indian Approach”, Bimabazaar.com Insurance Knowledge Portal, available at <http://www.bimabazaar.com/index.php/2013-04-05-07-10-11/86-fraud-in-motor-and-health-insurance-global-perspective-indian-approach>, accessed on 20 March 2017.
3. Anon (2016), “Harnessing the power of telecom data”, Hewlett Packard Enterprise Business White Paper, available at <https://h20195.www2.hpe.com/V2/getpdf.aspx/4AA6-4370ENW.pdf?ver=1.0>.
4. Abraham P, Pradhan M, Lakshminarayanan, Iyer G and Kumar U D (2016), “Customer Analytics at Bigbasket – Product Recommendations”, Indian Institute of Management Bangalore Case Study, IMB 573, available for download at: <https://cb.hbsp.harvard.edu/cbmp/product/IMB573-PDF-ENG>
5. Bhansali N, Rudravaram J, Grover S and Kumar U D (2016), “Customer Analytics at Flipkart”, IIM Bangalore Case IMB 555.
6. Brody H, Rip M R, Johansen P V, Paneth N, and Rachman S (2000), “Map Making and Myth Making in Broad Street: The London Cholera Epidemic, THE LANCET, 356(1), 64–68, 2000.
7. Cameron D and Jones I G (1982), “John Snow, the Broad Street Pump and the Model Epidemiology”, International Journal of Epidemiology, 12 (4), 393–396, 1983.
8. Duhigg C (2012), “The Power of Habit”, William Heinemann, London, 2012.
9. Coase R H (1937), “The Nature of the Firm”, *Economica*, 4, 386–405, 1937.
10. Coles P A, Lakhani K R and McAfee A P (2007), “Prediction Markets at Google”, *Harvard Business School Case* (Case Number 9-613-045).
11. Carneiro H A and Mylonakis E (2009), “Google Trends: A Web Based Tool for Real Time Surveillance of Disease Outbreaks”, 49(10), 1557–1564.

12. Davenport T H and Harris J G (2007), "Competing on Analytics – The New Science of Winning", *Harvard Business Publishing School Corporation*.
13. Davenport T H and Patil D J (2012), "Data Scientists: The Sexiest Job of the 21<sup>st</sup> Century", *Harvard Business Review*, 70–77, October 2012.
14. Davenport T H, Iansiti M and Serels A (2013), "Managing with Analytics at Procter and Gamble", *Harvard Business School Case* (Number 9-613-045).
15. Dinesh Kumar U, Shenoy S and Pandit A (2014), "Data Driven Decision Making", CII Report.
16. Fama E F (1980), "Problems and Theory of Firm", *Journal of Political Economy*, 88(2), 288–307, 1980.
17. Fernandes D, Puntoni S, van Osselaer S M J, and Cowley E (2013), "When and Why we Forget to Buy", *Journal of Consumer Psychology*, 26(3), 363–380, 2016.
18. Foer (J 2006), "The Kiss of Life", *Opinion Pages, New York Times*, February 14, 2006.
19. Ginsberg J, Mohebbi M H, Patel R S, Brammer L, Smolinski M S, and Brilliant L (2009), "Detecting Influenza Epidemics Using Search Engine Query Data", *Nature*, 457, 1012–1014.
20. Green K (2006) "The \$1 Million Netflix Challenge", *MIT Technology Review*, October 2006.
21. Haggstrom O (2007), "Problem Solving is often a matter of Cooking up an Appropriate Markov Chain", *Scandinavian Journal of Statistics*, 43(4), 768–780.
22. Hayes B (2013), "First Links in the Markov Chain", *American Scientist*, 101, 92–97.
23. Hays C L (2004), "What Wal-Mart Knows About Customers' Habits", *New York Times*, November 14, 2004.
24. Hemann C and Burbary K (2013), "Digital Marketing Analytics - Making Sense of Consumer Data in Digital World", Que Publishing, New York, USA.
25. Hopkins M S, LaValle S, Balboni F, Kruschwitz N, and Shockley R (2010), "10 Insights: A First look at the New Intelligence Enterprise Survey on Winning with Data", *MIT Sloan Management Review*, 52, 21–31.
26. Hu H and Jasper C R (2004), "Men and Women: A Comparison of Shopping Mall Behaviour", *Journal of Shopping Center Research*, 11(1–2), 113–131.
27. Kant G, Jacks M, and Aantjes C (2008), "Coca-Cola Enterprises Optimizes Vehicle Routes for Efficient Product Delivery", *Interfaces*, 38, 40–50.
28. Kent C (2015), "Why More Relationships End in the First Week of January?", *The Telegraph*, January 20, 2015.
29. Kumar V and Mirchandani R (2012), "Increasing the ROI of Social Marketing", *MIT Sloan Management Review*, 54, 55–61.
30. Lewis M (2003), "Moneyball: The Art of Winning an Unfair Game", W W Norton and Company, New York.
31. Leachman R C, Kang J, and Lin V (2002), "SLIM: Short Cycle Time and Low Inventory in Manufacturing at Samsung Electronics", *Interfaces*, 32, 61–67.
32. Li D K (2014), "Man uses First-Class Plane Ticket to Eat Free for a Year", *New York Post*, January 29, 2014.
33. Lino M, Kuczynski K, Rodriguez N, and Schap T (2017), "Expenditures on Children by Families, 2015", United States Department of Agriculture Report Number 1528–2015, January 2017.
34. MacKenzie I, Meyer C, and Noble S (2013), "How Retailers can keep up with Consumers", *McKinsey & Company Insights*, October 2013. Available at <http://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers> accessed on 20 March 2017.
35. Mahadevan B, Sivakumar S, Kumar D, and Ganeshram K (2013), "Redesigning Mid-day Meal Logistics for the Akshaya Patra Foundation: OR at work in Feeding Hungry School Children", *Interfaces*, 43(6), 530–546.
36. Martin J A, Hamilton B E, Osterman M J K, Driscoll A K, and Mathews T J (2017), "National Vital Statistics Reports Births Final Data for 2015", 66(1).
37. Mayer-Schonberger V and Cukier K (2013), "Big Data – A Revolution that will Transform how we Live, Work and Think", John Murray, London.
38. Carbonell J G, Michalski R S, and Mitchell T M, "An overview of Machine Learning in Machine Learning – An Artificial Intelligence Approach (Eds) R S Michalski, J G Carbonell and T M Mitchell, Springer, New York, 1983.
39. Michalos A C (1970), "The costs of decision making", *Public Choice*, 9, 39–51.
40. Mitchell T M (2006), "The Discipline of Machine Learning", Carnie Mellon University Report CMU-ML-06-108, available at <http://www.cs.cmu.edu/~tom/pubs/MachineLearning.pdf>, accessed on 10 April 2017.
41. Ransbotham S and Kiron D (2017), "Analytics as Source of Innovation", *MIT Sloan Management Review*, 1–17, Spring 2017.
42. Savant M V (1990), "Ask Marilyn", *Parade Magazine*, Page 16, 9 September 1990.

43. Scott N (2013), "Big Data: What's the Big Deal?", Director, 61–65.
44. Selvin S (1975), "A Problem in Probability", Letters to the Editor – *The American Statistician*, 29(1), 67–71.
45. Siegel E (2013), "Predictive Analytics: The Power to Predict who will Click, Buy, Lie or Die", John Wiley and Sons, Hoboken, NJ.
46. Simon H (1972), "Theories of Bounded Rationality", in *Decisions and Organizations by C B McGuire and R Radner (Eds)*, North Holland Publishing Company.
47. Sirkin H L, Keenan P, and Jackson A (2009), "The Hard Side of Change Management", *Harvard Business Review*, 108–118.
48. Snow S J (2002), "Commentary: Sutherland, Snow and Water: The Transmission of Cholera in the nineteenth Century", *International Journal of Epidemiology*, 31, 908–911.
49. Srujana H M, Saranga H, and Kumar UD "Era of Quality at Akshaya Patra", IIM Bangalore Case IMB 493.
50. Stelzner M A (2013), "Social Media Marketing Industry Report – How Marketers Using Social Media to Grow Their Businesses", *Social Media Examiner Business Report*.
51. Suhruta K, Makija K, and Kumar U D (2013), "1920 Evil Returns – Bollywood and Social Media Marketing", IIM Bangalore Case (Case Number IMB 437).
52. Snyder B (2013), "9 Facts About Walmart that will Surprise You", *Fortune*, June 06, 2015.
53. Tandon R, Chakraborty A, Srinivasan G, Shroff M, Abdulla A, Shamsundar B, Sinha R, Subramaniam S, Hill D, and Dhore P (2013), "Hewlett Packard: Delivering Profitable Growth for HPDirect.Com using Operational Research", *Interfaces*, 43(1), 48–61.
54. Tufte E (2001), "Visual Display of Quantitative Information", Graphics Press, Connecticut.
55. Underhill P (2009), "Why We Busy – The Science of Shopping", Simon & Schuster Paperbacks, New York.
56. Whitelocks S (2013), "Having nightmares about your husband cheating? It may might be true. New research finds Dream can predict future relationship behaviour", Daily Mail, 16 May 2013.



# Descriptive Analytics

2

“The Purpose of Visualization is Insight – Not Pictures”.

—Ben Shneiderman

## LEARNING OBJECTIVES

- LO 2-1** Understand the basic concepts in descriptive analytics and how it is used in data-driven decision making.
- LO 2-2** Learn different variable types such as qualitative and quantitative along with scales of measurement such as nominal, ordinal, interval and ratio.
- LO 2-3** Understand data types such as cross-sectional data, time series data and panel data.
- LO 2-4** Understand the difference between population and sample and gain insights through fundamental concepts in statistics such as measures of central tendency, measures of variability and measures of shape.
- LO 2-5** Learn data visualization and various types of visual charts.
- LO 2-6** Understand the application of descriptive analytics in decision making.

## ESSENCE OF DESCRIPTIVE ANALYTICS

Descriptive analytics is about finding “what has happened” by summarizing the data using innovative methods and analysing the past data using simple queries. Analysing past data can provide insights that can assist organizations to take appropriate decisions. Consider the Walmart example discussed in Chapter 1, where they found that during the hurricane season the demand for strawberry pop-tart increased seven times the normal season; this is a very good example for application of descriptive statistics. Based on this insight, Walmart ensured that there is enough stock of strawberry pop-tarts in the stores during a hurricane season. John Snow’s spot map on Cholera outbreak in London and his final hypothesis that Cholera is water-borne disease is another classic example of application of descriptive analytics through data visualization. There are many such examples where simple analysis of the past data has revealed interesting facts such as difference in shopping behaviour of men and women, relationship freeze, etc. Descriptive analytics involves data summarization – using techniques such as pivot tables, descriptive statistics and data visualization that can be used for analysing past data to gain insights and hidden patterns.

**IMPORTANT**

*Descriptive analytics is the starting point of analytics based solution to problems. It helps to understand the data and provide directions for predictive and prescriptive analytics. Business Intelligence (BI), which largely involves creating reports and business dashboard that lead to actionable insights, is essentially a descriptive analytics exercise.*

## 2.1 | INTRODUCTION TO DESCRIPTIVE ANALYTICS

Descriptive analytics is the science of describing past data and thus capturing “what happened” in a given context. Primary objective of descriptive analytics is simple comprehension of data using data summarization, basic statistical measures and visualization. Various tools and techniques are used in describing the data. Descriptive statistics such as measures of central tendency, measures of variation and measures of shape can provide useful insights. Many different plots such as histogram, bar chart, pie-chart, box-plot, scatter plot and tree diagram can provide insights about past data and subsequently assist with further analysis by generating new hypotheses.

Descriptive analytics is an important part of reporting across several industries which enables top management to monitor key performance indicators and take decisions. Most companies generate reports and dashboards at regular intervals as part of business intelligence (BI) to communicate various aspects of the business to the top management, stakeholders, and the external world. Business reports include descriptive analytics in the form of tables, charts, and innovative diagrams such as Treemap. With the advent of mobile technology, many real-time reports are generated and are accessed by the top management in their mobile handsets enabling them to take quick actions if necessary. For example, a retailer such as Bigbazaar or Reliance retail in India may like to know the top 5 (in terms of revenue generated) products that are sold by region, by city, by store, etc. Such information would assist the management to plan their inventory, shelf space, pricing, etc. They can also monitor trend in revenue generated at regional, city, and store levels over the past several periods. Several companies use dashboards and scorecards to communicate KPIs that are relevant to them; one of the primary applications of descriptive analytics is designing effective dashboards and scorecards.

## 2.2 | DATA TYPES AND SCALES

Data is classified into different categories based on data structure and scale of measurement of the variables.

### 2.2.1 | Structured and Unstructured Data

Data at a macro-level can be classified as structured and unstructured data. Structured data means that the data is described in a matrix form with labelled rows and columns. Any data that is not originally in the matrix form with rows and columns is an unstructured data. For example, e-mails, click streams, textual data, images (photos and images generated by medical devices), log data, and videos. Machine-generated data such as images generated by satellite, magnetic resonance imaging (MRI), electrocardiogram (ECG) and thermography are few examples of unstructured data. There is an increasing trend in

the generation of unstructured data due to social media platforms such as Facebook and YouTube and analysis of unstructured data is important for effective management. Internet of things (IoT) is another source unstructured data.

The importance of unstructured data in decision making has increased many folds in the recent past due to its applications to different sectors of the industry. For example, analysing social media data is important for companies to understand the sentiments expressed by the customers about their products/services and take necessary remedial measures. Significant proportion of social media data is natural language (text) apart from images and videos. Apart from social media, machine-generated data are usually unstructured (e.g. data generated from medical devices such as ECG, MRI, etc.). High percentage of Big Data problems constitute unstructured data. One of the main challenges in analysing unstructured data is in the conversion of unstructured data to structured data, which then enables model development. Examples of structured and unstructured data are shown in Tables 2.1 and 2.2.

The data in Table 2.2 is a clickstream data (search behaviour of an internet user that captures the websites visited by the user). Clickstream data is useful for understanding the behaviour of internet users. Based on their surfing (internet browsing) behaviour, individuals are targeted with advertisement for products and services. The unstructured data as shown in Table 2.2 does not have matrix structure as in the case of structured data in Table 2.1. Before any analytics model can be built, unstructured data has to be converted into a structured data.

**TABLE 2.1** Structured data consisting of nominal and ratio scales

No.	Gender	Age	Percentage SSC	Board SSC	Percentage HSC	Percentage Degree	Salary
1	M	23	62	Others	88	52	270000
2	M	21	76.33	ICSE	75.33	75.48	220000
3	M	22	72	Others	78	66.63	240000
4	M	22	60	CBSE	63	58	250000
5	M	22	61	CBSE	55	54	180000
6	M	23	55	ICSE	64	50	300000
7	F	24	70	Others	54	65	240000
8	M	22	68	ICSE	77	72.5	235000
9	M	24	82.8	CBSE	70.6	69.3	425000
10	F	23	59	CBSE	74	59	240000

**TABLE 2.2** Unstructured data (sample clickstream data)

<https://en.wikipedia.org/wiki/Clickstream>

<http://hortonworks.com/hadoop-tutorial/how-to-visualize-website-clickstream-data/>

<http://searchcrm.techtarget.com/definition/clickstream-analysis>

<https://www.qubole.com/blog/big-data/clickstream-data-analysis/>

## 2.2.2 | Cross-sectional, Time Series, and Panel Data

Another important classification of data is based on the type of data collected. Based on the type of data collected, the data is grouped into the following three classes:

- Cross-Sectional Data:** A data collected on many variables of interest at the same time or duration of time is called cross-sectional data. For example, consider data on movies such as budget, box-office collection, actors, directors, genre of the movie during year 2017.
- Time Series Data:** A data collected for a single variable such as demand for smartphones collected over several time intervals (weekly, monthly, etc.) is called a time series data.
- Panel Data:** Data collected on several variables (multiple dimensions) over several time intervals is called panel data (also known as longitudinal data). Example of a panel data is data collected on variables such as gross domestic product (GDP), Gini index, and unemployment rate for several countries over several years.

## 2.3 | TYPES OF DATA MEASUREMENT SCALES

Structured data can be either numeric or alpha numeric and may follow different scales of measurement (level of measurement). It is important to understand the type of variables within the data with respect to the measurement scale since the model specification while building analytics models such as regression may depend on the scale of measurement.

### 2.3.1 | Nominal Scale (Qualitative Data)

Nominal scale refers to variables that are basically names (qualitative data) and also known as categorical variables. For example, variables such as marital status (single, married, divorced) and industry type (manufacturing, healthcare, banking and finance) fall under nominal scale. During data collection, it is usual to assign a numerical code to represent a nominal variable. For example, the data collector may have used number 1 to represent singles, 2 for married, and 3 for divorced category for categorical variable marital status. The codes 1, 2, and 3 used here do not have any value attached to them. That is, basic mathematical operations are meaningless in a nominal scale (e.g., subtraction: married – unmarried or ratio: married/unmarried are meaningless). While developing statistical models, nominal scale data are usually transformed before building the model. For example, when developing a regression model, categorical variables are converted using dummy variables before building the regression model (is discussed in Chapter 10).

### 2.3.2 | Ordinal Scale

Ordinal scale is a variable in which the value of the data is captured from an ordered set, which is recorded in the order of magnitude. For example, in many survey data, Likert scale is used. Likert scale is finite (usually a 5 point scale) and the data collector would have defined the order of preference. For example, assume that a feedback is collected on a training program using 5-point Likert scale in which 1 = Poor, 2 = Fair, 3 = Good, 4 = Very Good, and 5 = Excellent. In this case, we know that

5 is better than 4 and 4 is better than 3; however, the difference 5 – 4 (Excellent – Very Good) is meaningless.

### 2.3.3 | Interval Scale

Interval scale corresponds to a variable in which the value is chosen from an interval set. Variable such as temperature measured in centigrade ( $^{\circ}\text{C}$ ) or intelligence quotient (IQ) score are examples of interval scale. In interval scale, the ratios do not make sense. For example,  $40^{\circ}\text{C}$  is not twice hot as  $20^{\circ}\text{C}$ . Similarly, a person with an IQ score of 160 is not twice smarter than a person with an IQ score of 80. However,  $40^{\circ}\text{C}$  is  $20^{\circ}\text{C}$  more than  $20^{\circ}\text{C}$ , IQ score of 160 is 80 more than an IQ score of 80. In an interval scale, the reference is fixed arbitrarily, for example  $0^{\circ}\text{C}$  is fixed based on the freezing point of water.

### 2.3.4 | Ratio Scale

Any variable for which the ratios can be computed and are meaningful is called ratio scale. Most variables come under this type; for example: demand for a product, market share of a brand, sales, salary, and so on. If Ms Hawai Sundari's salary is 40,000 per month and Ms Dawai Sundari's salary is 90,000 per month then we can interpret that Dawai Sundari earns 2.25 times the salary of Hawai Sundari.

## 2.4 | POPULATION AND SAMPLE

**Population** is the set of all possible observations (often called cases, records, subjects or data points) for a given context of the problem. The size of the population can be very large in many cases. For example, in 2014, close to 834.08 million people were eligible to vote in the Indian general elections (Source: Election Commission of India). Thus, the population size of the voters in 2014 was 834.08 million which included all eligible voters. During every election, media and other organizations collect data to predict likely winner of election through opinion polls (and they rarely get it right due to complexities associated with collecting right sample). It is very difficult (also practically impossible) to collect data from all 834.08 million eligible voters about their choice of candidate, so the opinion polls are based on opinion expressed by a subset of voters called **sample**.

Population (also known as universal set) is the set of all possible data for a given context whereas sample is the subset taken from a population. In many analytical problems, we make inference about the population based on the sample data. There are many challenges in sampling (process of selecting an observation from the population). An incorrect sample may result in bias and incorrect inference about the population. Sampling is discussed in detail in Chapter 4.

## 2.5 | MEASURES OF CENTRAL TENDENCY

Measures of central tendency are the measures that are used for describing the data using a single value. **Mean, median** and **mode** are the three measures of central tendency and are frequently used to compare different data sets. Measures of central tendency help users to summarize and comprehend the data.

### 2.5.1 | Mean (or Average) Value

Mean is the arithmetical average value of the data and is one of the most frequently used measures of central tendency. Assume that the data has  $n$  observations in a sample, and let  $X_i$  be the value of the  $i^{\text{th}}$  observation. Then the mean is given by

$$\text{Mean} = \bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} = \sum_{i=1}^n \frac{X_i}{n} \quad (2.1)$$

Symbol  $\bar{X}$  is frequently used to represent the estimated value of the mean from a sample. If the entire population is available and if we calculate mean based on the entire population, then we get the population mean which is denoted by  $\mu$ . Among the measures of central tendency, mean is the most frequently used measure since it uses all the observations (all  $X_i$  values) in the data set (either sample or population) to calculate the mean value. Table 2.1 has the salary of graduating students from a business school; the average salary is given by

$$\bar{X} = \frac{(270 + 220 + 240 + 250 + 180 + 300 + 240 + 235 + 425 + 240) \times 1000}{10} = 260000$$

The average (or mean) salary is 260000. Note that the average value need not be a part of the data set, that is, none of the graduating student's salary is 260000. In Microsoft Excel, function 'Average(array)' can be used for calculating the mean value of the data. Mean can be interpreted as the centre of gravity of the distribution of the data. An important property of mean is that the summation of deviation of observations from the mean is zero, that is

$$\sum_{i=1}^n (X_i - \bar{X}) = 0$$

Associated with the mean is a phenomenon often called "wisdom of crowd", according to which the collective wisdom of people is better than any individual person's knowledge. For example, in 1906, Francis Galton attended a contest in Plymouth, UK in which the villagers were asked to guess the weight of an Ox, the one who guessed the closest won the prize. Around 800 villagers participated in the contest. Francis Galton found that the average of all the weights entered was very close to the actual weight. In fact, the difference was less than a pound. Also, the average turned to be better than the guess by the winner of the contest (Surowiecki, 2004).

One should be careful about taking decisions based on the mean value of the data. There is a famous joke in statistics which says that, "*if someone's head is in freezer and leg is in the oven, the average body temperature would be fine, but the person may not be alive*". Making decisions solely based on mean value is not advisable. In capital asset procurement such as procurement of fighter aircraft and weapons, defence services across the world use mean time between failures (MTBF) as one of the measures of system reliability (performance). However, MTBF (which is the mean value of the time between failure data) in itself is not a useful measure to assess the reliability of the asset and not very useful in taking operational decisions. It has to be used along with other measures and measures of variability for better understanding of the data. Another issue with mean is, it is affected significantly by presence of

outliers. That is, presence of an outlier can change the mean value significantly. If the data is captured in frequencies, then Eq. (2.2) can be used for calculating the average:

$$\bar{X} = \sum_{i=1}^n \frac{f_i X_i}{f_i} \quad (2.2)$$

The frequency of age of students in Table 2.1 is given below:

Age	21	22	23	24
Frequency	1	4	3	2

The average age of students using Eq. (2.2) is given by

$$\bar{X} = \frac{1 \times 21 + 4 \times 22 + 3 \times 23 + 2 \times 24}{1 + 4 + 3 + 2} = 22.6$$

## 2.5.2 | Median (or Mid) Value

Median is the value that divides the data into two equal parts, that is, the proportion of observations below median and above median will be 50%. Easiest way to find the median value is by arranging the data in the increasing order and the median is the value at position  $(n + 1)/2$  when  $n$  is odd. When  $n$  is even, the median is the average value of  $(n/2)^{\text{th}}$  and  $(n + 2)/2^{\text{th}}$  observation after arranging the data in the increasing order.

Consider the example of a bank. The number of deposits in a branch of a bank in a week is shown in Table 2.3.

**TABLE 2.3** Number of daily deposits in a Bank

Day	1	2	3	4	5	6	7
Number of Deposits	245	326	180	226	445	319	260

The ascending order of the data in Table 2.3 is given by

180, 226, 245, 260, 319, 326 and 445

Now  $(n + 1)/2 = (8/2) = 4$ . Thus the median is the 4<sup>th</sup> value in the data after arranging them in the increasing order; in this case it is 260. There are equal numbers of observation below and above 260. In Microsoft Excel, the function ‘Median(array)’ can be used for calculating the median of a data set.

Another example is the salary in Table 2.1 that can be arranged as follows:

180000, 220000, 235000, 240000, 240000, 240000, 250000, 270000, 300000, 425000

The 5<sup>th</sup> and 6<sup>th</sup> observations are 240000 and 240000 and the average is 240000. Thus, the median salary for the data in Table 2.1 is 240000. Median is much more stable than the mean value, that is adding a new observation may not change the median significantly. However, the drawback of median is that it is not calculated using the entire data like in the case of mean. We are simply looking for the midpoint instead of using the actual values of the data.

### 2.5.3 | Mode

Mode is the most frequently occurring value in the data set. For example, in the data ‘salary’ in Table 2.1, the value 240000 is appearing three times and is the mode since all other values are observed only once. In Microsoft Excel, the function ‘Mode(array)’ can be used for calculating mode. Mode is the only measure of central tendency which is valid for qualitative (nominal) data since the mean and median for nominal data are meaningless. For example, assume that a customer data with a retailer has the marital status of customer, namely, (a) Married, (b) Unmarried, (c) Divorced Male, and (d) Divorced Female. Mean and median are meaningless when we try to use them on a qualitative data such as marital status. On the other hand, mode will capture the customer type in terms of marital status that occurs most frequently in the database. In the bar chart (and histogram), mode is the tallest column. It is possible that a data set may not have any mode at all. For example, if each value in the data set appears only once, then there is no mode in the data set.

## 2.6 | PERCENTILE, DECILE, AND QUARTILE

Percentile, decile and quartile are frequently used to identify the position of the observation in the data set. Percentile score is frequently used in education to identify the position of a student in the group. Another frequent application of percentile is the percentile life used in asset management. Percentile, denoted as  $P_x$ , is the value of the data at which  $x$  percentage of the data lie below that value. For example,  $P_{10}$  denotes the value below which 10 percentage of the data lies. To find  $P_x$ , we have to arrange the data in the increasing order and the value of  $P_x$  is the position in the data calculated using Eq. (2.3):

$$\text{Position corresponding to } P_x \approx \frac{x(n+1)}{100} \quad (2.3)$$

where  $n$  is the number of observations in the data. Note that the value obtained from Eq. (2.3) can be non-integer, in which case we can either round it to the nearest integer or use an approximation which will be explained in Example 2.1. **Decile** corresponds to special values of percentile that divide the data into 10 equal parts. First decile contains first 10% of the data and second decile contains first 20% of the data and so on. Similarly, **Quartile** divides the data into 4 equal parts. The first quartile ( $Q_1$ ) contains first 25% of the data,  $Q_2$  contains 50% of the data and is also the median. Quartile 3 ( $Q_3$ ) accounts for 75% of the data. In Microsoft Excel, the function ‘Percentile(array, k)’ provides  $P_x$  value. That is, Percentile(array, 0.1) will give 10<sup>th</sup> percentile.

#### EXAMPLE 2.1

Time between failures (in hours) of a wire cutter used in a cookie manufacturing oven is given in Table 2.4. The function of the wire-cut is to cut the dough into cookies of desired size.

**TABLE 2.4** Time between failures of wire-cut (in hours)

2	22	32	39	46	56	76	79	88	93
3	24	33	44	46	66	77	79	89	99
5	24	34	45	47	67	77	86	89	99
9	26	37	45	55	67	78	86	89	99
21	31	39	46	56	75	78	87	90	102

- (a) Calculate the mean, median, and mode of time between failures of wire-cuts.
- (b) The company would like to know by what time 10% (ten percentile or  $P_{10}$ ) and 90% (ninety percentile or  $P_{90}$ ) of the wire-cuts will fail?
- (c) Calculate the values of  $P_{25}$  and  $P_{75}$ .

**Solution:**

- (a) Mean = 57.64, median = 56, and mode = 46, 89 and 99.
- (b) Note that the data in Table 2.4 is arranged in increasing order in columns. The position of  $P_{10} = 10 \times (51)/100 = 5.1$ . We can round off 5.1 to its nearest integer which is 5. The corresponding value from table is 21 (10 percentage of observations in Table 2.4 have a value of less than or equal to 21). That is, by 21 hours, 10% of the wire-cuts will fail. In asset management (and reliability theory), this value is called  $P_{10}$  life.

Instead of rounding the value obtained from Eq. (2.3), we can use the following approximation:

$$\text{Position corresponding to } P_{10} = 10 \times (51)/100 = 5.1$$

Value at 5<sup>th</sup> position is 21. Value at position 5.1 is approximated as

$$21 + 0.1 \times (\text{value at 6}^{\text{th}} \text{ position} - \text{value at 5}^{\text{th}} \text{ position}) = 21 + 0.1(1) = 21.1$$

$$\text{Position corresponding to } P_{90} = 90 \times 51/100 = 45.9$$

The value at position 45 is 90 and the value at position 45.9 is

$$90 + 0.9 \times (\text{value at 46}^{\text{th}} \text{ position} - \text{value at 45}^{\text{th}} \text{ position}) = 90 + 0.9 \times (3) = 92.7$$

That is, 90% of the wire-cuts will fail by 92.7 hours.

- (c) Position corresponding to  $P_{25}$  (1<sup>st</sup> Quartile or  $Q_1$ ) =  $25 \times 51/100 = 12.75$   
Value at 12<sup>th</sup> position is 33, so  
$$P_{25} = 33 + 0.75 (\text{value at 13}^{\text{th}} \text{ position} - \text{value at 12}^{\text{th}} \text{ position}) = 33 + 0.75 (1) = 33.75$$

$$\text{Position corresponding to } P_{75} \text{ (3}^{\text{rd}} \text{ Quartile or } Q_3\text{)} = 75 \times 51/100 = 38.25$$

Value at 38<sup>th</sup> position is 86, so

$$P_{75} = 86 + 0.25 (\text{value at 39}^{\text{th}} \text{ position} - \text{value at 38}^{\text{th}} \text{ position}) = 86 + 0.25 (0) = 86$$

## 2.7 | MEASURES OF VARIATION

One of the primary objectives of analytics is to understand the variability in the data. Predictive analytics techniques such as regression attempt to explain variation in the outcome variable ( $Y$ ) using predictor variables ( $X$ ). Variability in the data is measured using the following measures:

1. Range
2. Inter-Quartile Distance (IQD)
3. Variance
4. Standard Deviation

Let us discuss each of them in detail.

### 2.7.1 | Range

Range is the difference between maximum and minimum value of the data. It captures the data spread. In the data in Table 2.4, the range =  $102 - 2 = 100$ .

### 2.7.2 | Inter-Quartile Distance (IQD)

Inter-quartile distance (IQD), also called inter-quartile range (IQR), is a measure of the distance between Quartile 1 ( $Q_1$ ) and Quartile 3 ( $Q_3$ ). For the data in Table 2.4, we calculated  $Q_1$  as 33.75 and  $Q_3$  as 86. Thus the  $\text{IQD} = 86 - 33.75 = 52.25$ . IQD is a useful measure for identifying outliers in the data. Outlier is an observation which is far away (on either side) from the mean value of the data. Values of data below  $Q_1 - 1.5 \text{ IQD}$  and above  $Q_3 + 1.5 \text{ IQD}$  are classified as outliers.

For the data in Table 2.4

$$\begin{aligned} Q_1 - 1.5 \text{ IQD} &= 33.75 - 1.5 \times 52.25 = -44.625 \\ Q_3 + 1.5 \text{ IQD} &= 86 + 1.5 \times 52.25 = 164.375 \end{aligned}$$

In Table 2.4, there are no values either below  $-44.625$  or above  $164.375$ , thus there are no outliers. Note that IQD is one of the approaches used for identifying outliers; we will discuss other approaches that are used for identifying outliers in Chapters 9 and 10. Also, using IQD for identifying outliers is appropriate only in the case of univariate data (data with one dimension). In the case of multivariate data, we use distance measures such as Mahalanobis distance to identify outliers (discussed in Chapters 9 and 10).

### 2.7.3 | Variance and Standard Deviation

Variance is a measure of variability in the data from the mean value. Variance for population,  $\sigma^2$ , is calculated using

$$\text{Variance} = \sigma^2 = \sum_{i=1}^n \frac{(X_i - \mu)^2}{n} \quad (2.4)$$

Note that, in Eq. (2.4), deviation from mean is squared since sum of deviations from mean will always add up to zero. The variance for the data in Table 2.4 is 818.0304 [using Eq. (2.4)]. In case of a sample, the Sample Variance ( $S^2$ ) is calculated using

$$S^2 = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n-1} \quad (2.5)$$

While calculating sample variance  $S^2$ , the sum of squared deviation  $\sum_{i=1}^n (X_i - \bar{X})^2$  is divided by  $(n - 1)$ . This is known as Bessel's correction. For the data in Table 2.4, the sample standard variance is 834.7249. Microsoft Excel functions Var.P(array) and Var.S(array) are used for calculating population variance and sample variance, respectively. The population standard deviation ( $\sigma$ ) and sample standard deviation ( $S$ ) are given by

$$\sigma = \sqrt{\sum_{i=1}^n \frac{(X_i - \mu)^2}{n}} \quad (2.6)$$

For the data in Table 2.4, the standard deviation obtained using the Eq. (2.6) is 28.6012.

$$S = \sqrt{\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n-1}} \quad (2.7)$$

For the data in Table 2.4, the standard deviation obtained using the Eq. (2.7) is 28.8916. In Microsoft Excel, functions Stdev.P(array) and Stdev.S(array) are used for calculating population standard deviation and sample standard deviation respectively. There are two arguments for dividing the sum of squared deviations from mean by  $(n - 1)$  instead of  $n$  in Eqs. (2.5) and (2.7). One argument is that, when we take a sample and estimate the mean from the sample  $\bar{X}$ , we tend to underestimate the sum of squared deviations from the mean. For example, take a sample consisting of first 5 (first column) and last 5 (last column) observations from Table 2.4. The sample is given in Table 2.5.

**TABLE 2.5** Sample of 10 observations from Table 2.4

2	3	5	9	21	93	99	99	99	102
---	---	---	---	----	----	----	----	----	-----

The mean  $\bar{X}$  for the sample in Table 2.5 is 53.2 and standard deviation [using Eq. (2.7)] is 47.9740. When we estimate the numerator,  $(X_i - \mu)^2$ , in Eq. (2.4) using  $\bar{X}$ , instead of  $\mu$ , we will underestimate  $(X_i - \mu)^2$  resulting in underestimation of standard deviation. The calculations of  $(X_i - \bar{X})^2$  and  $(X_i - \mu)^2$  for the sample in Table 2.5 are shown in Table 2.6.

**TABLE 2.6** Underestimation of standard deviation in sample

Data	Standard deviation (using sample mean 53.2)	Standard deviation (using population mean 57.64)
2	2621.44	3095.81
3	2520.04	2985.53
5	2323.24	2770.97
9	1953.64	2365.85
21	1036.84	1342.49
93	1584.04	1250.33
99	2097.64	1710.65
99	2097.64	1710.65
99	2097.64	1710.65
102	2381.44	1967.81
Sample Mean = 53.2	$\sum (X_i - \bar{X})^2 = 20713.60$	$\sum (X_i - \mu)^2 = 20910.74$

In Table 2.6, we can see that the numerator in Eq. (2.4) is underestimated (20713.60) when we use the sample average against population average (20910.74). This will result in underestimation of the standard deviation, a phenomenon called **downward bias**. To overcome this bias, we divide  $\sum (X_i - \bar{X})^2$  with  $(n - 1)$  instead of  $n$ .

Another argument of using Eq. (2.5) is through the concept of **degrees of freedom**. The following two definitions are used for degrees of freedom (Pandey and Bright, 2008):

1. Degrees of freedom is equal to the number of independent variables in the model (Trochim, 2005). For example, we can create any sample of size  $n$  with mean value of  $\bar{X}$  by randomly selecting  $(n - 1)$  values. We need to fix just one out of  $n$  values. Thus the number of independent variables in this case is  $(n - 1)$ .
2. Degrees of freedom is defined as the difference between the number of observations in the sample and number of parameters estimated (Walker 1940, Toothaker and Miller, 1996). If there are  $n$  observations in the sample and  $k$  parameters are estimated from the sample, then the degrees of freedom is  $(n - k)$ . While using Eq. (2.5) or Eq. (2.7), the value of  $\bar{X}$  is estimated from the sample. Thus the degrees of freedom is  $(n - 1)$ .

Whenever we estimate a parameter from a sample, we lose a degree of freedom. While estimating standard deviation from a sample, we tend to underestimate since mean is also estimated from the sample itself. The downward bias is addressed by dividing the sum of squared deviation from mean with  $(n - 1)$  instead of  $n$ .

## 2.7.4 | Chebyshev's Theorem

Chebyshev's theorem (also known as Chebyshev's inequality) is an empirical rule that allows us to predict proportion of observations that is likely to lie between an interval defined using mean and standard deviation. Probability of finding a randomly selected value in an interval defined by  $\mu \pm k\sigma$  is  $\geq 1 - \frac{1}{k^2}$ , that is

$$P(\mu - k\sigma \leq X \leq \mu + k\sigma) \geq 1 - \frac{1}{k^2} \quad (2.8)$$

Equation (2.8) is useful when the value of  $k > 1$ , otherwise it gives a trivial solution.

### EXAMPLE 2.2

Amount spent per month by a segment of credit card users of a bank has a mean value of 12000 and standard deviation of 2000. Calculate the proportion of customers who are spending between 8000 and 16000.

**Solution:**

$$P(8000 \leq X \leq 16000) = P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \geq 1 - \frac{1}{2^2} = 0.75$$

That is, the proportion of customers spending between 8000 and 16000 is at least 0.75 (or 75%)

## 2.8 | MEASURES OF SHAPE – SKEWNESS AND KURTOSIS

Skewness is a measure of symmetry or lack of symmetry. A data set is symmetrical when the proportion of data at equal distance (measured in terms of standard deviation) from mean (or median) is equal. That is, the proportion of data between  $\mu$  and  $\mu - k\sigma$  is same as  $\mu$  and  $\mu + k\sigma$ , where  $k$  is some positive constant. This implies that the distribution (or proportion) of the data on either side of mean (and median) is same. Measure of skewness can be used to identify whether the distribution is left skewed (longer tail on left side of the distribution) or right skewed (longer tail on the right side of the distribution). There are many different approaches to measuring skewness. **Pearson's moment coefficient of skewness** for a data set with  $n$  observations is given by

$$g_1 = \frac{\sum_{i=1}^n (X_i - \mu)^3 / n}{\sigma^3} \quad (2.9)$$

The value of  $g_1$  will be close to 0 when the data is symmetrical. A positive value of  $g_1$  indicates a positive skewness and a negative value indicates negative skewness. The formula in Eq. (2.9) is adjusted for sample size when skewness is calculated from a sample. The following formula is used usually for a sample with  $n$  observations (Joanes and Gill, 1998):

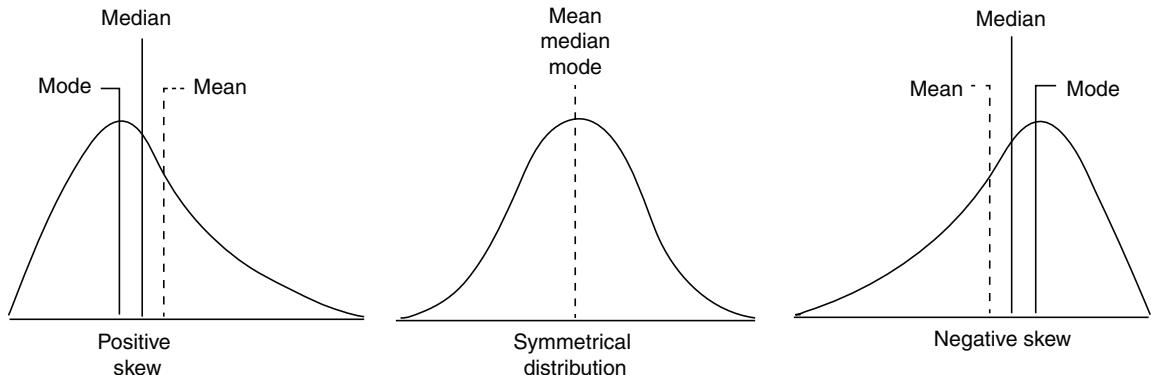


FIGURE 2.1 Skewness.

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1 \quad (2.10)$$

The value of  $\frac{\sqrt{n(n-1)}}{n-2}$  will converge to 1 as the value of  $n$  increases. For the data in Table 2.4, the

value of  $G_1$  is  $-0.232$ . Since the value of  $G_1$  is negative, we can conclude that the data is left skewed. In Microsoft Excel, function 'SKEW(array)' can be used for calculating the value of skewness ( $G_1$ ) calculated from a sample. In Figure 2.1, the positive skewed (right tailed), normal, and negative skewed (left tailed) distributions are shown.

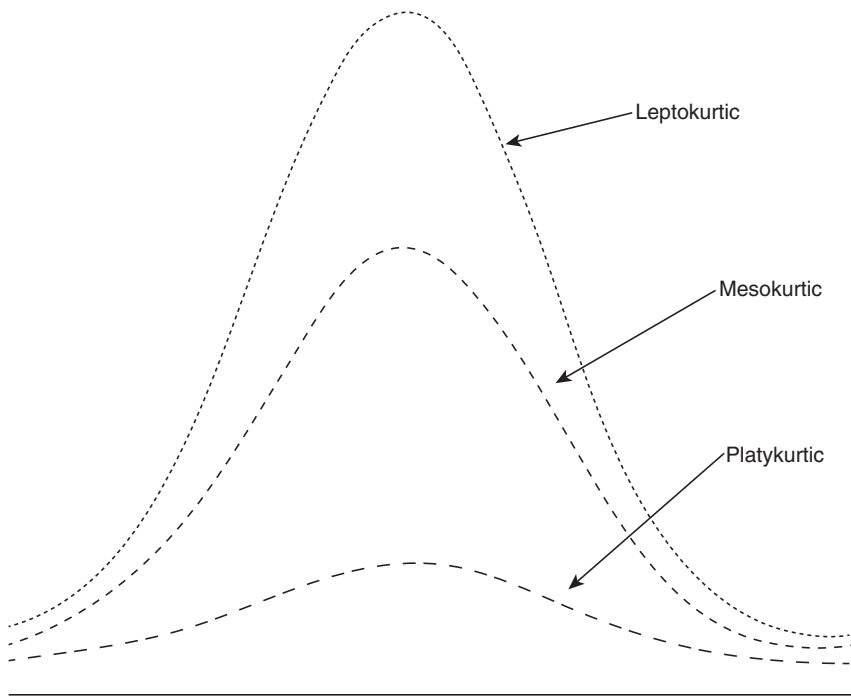
**Kurtosis** is another measure of shape, aimed at shape of the tail, that is, whether the tail of the data distribution is heavy or light. Kurtosis is measured using the following equation:

$$\text{Kurtosis} = \frac{\sum_{i=1}^n (X_i - \bar{X})^4 / n}{\sigma^4} \quad (2.11)$$

Kurtosis value of less than 3 is called **platykurtic distribution** and greater than 3 is called **leptokurtic distribution**. The kurtosis value of 3 indicates standard normal distribution (also called **mesokurtic**). The excess kurtosis is a measure that captures deviation from kurtosis of a normal distribution and is given by

$$\text{Excess Kurtosis} = \frac{\sum_{i=1}^n (X_i - \bar{X})^4 / n}{\sigma^4} - 3 \quad (2.12)$$

For the data in Table 2.4, excess Kurtosis =  $-1.0968$  (that is kurtosis is  $1.9032$ ). Figure 2.2 shows shapes of platykurtic, mesokurtic, and leptokurtic distributions. In Microsoft Excel, 'KURT(array)' can be used for calculating the excess kurtosis.



**FIGURE 2.2** Leptokurtic, mesokurtic, and platykurtic distributions.

## 2.9 | DATA VISUALIZATION

Data visualization is an integral part of descriptive analytics and it assists decision makers with useful insights. There are many useful charts such as histogram, bar chart, pie-chart, box-plot that would assist data scientist with visualization of the data. In the recent years, tree maps and sunburst maps are very popular among analytics experts, which can create hierarchical visuals of data. It is always advisable to start an analytics project with data visualization.

### 2.9.1 | Histogram

Histogram is the visual representation of the data which can be used to assess the probability distribution (frequency distribution) of the data. It is a frequency distribution of data arranged in consecutive and non-overlapping intervals. Histograms are created for continuous (numerical) data. The following steps are used in constructing histograms:

1. Divide the data into finite number of non-overlapping and consecutive bins (intervals). The total number of bins to be used can be calculated using Eqs. (2.13) or (2.14).
2. Count the number of observations from the data that fall under each bin (interval).
3. Create a frequency distribution (bin in the horizontal axis and frequency in the vertical axis) using the information obtained in steps 1 and 2.

Histogram is very useful since it assists data scientist to identify the following:

1. The shape of the distribution and to assess the probability distribution of the data.
2. Measures of central tendency such as median and mode.
3. Measures of variability such as spread.
4. Measure of shape such as skewness.

Histograms are also useful in identifying the presence of outliers. One of the first steps in constructing histogram is identifying the number of bins. There are many different formulas used in literature and one of the simplest formula is

$$\text{Number of bins, } N = \frac{X_{\max} - X_{\min}}{W} \quad (2.13)$$

Here  $X_{\max}$  and  $X_{\min}$  are the maximum and minimum values of the data and  $W$  is desired the width of the bin (interval). Intervals in histograms are usually of equal size. Sturges (1926) proposed the following formula for calculating the number of bins:

$$\text{Number of bins, } N = 1 + 3.322 \log_{10}(n) \quad (2.14)$$

where  $n$  is the total number of observations in the data set. Figures 2.3 and 2.4 show the histogram of Bollywood movie budget in crores of rupees (1 crore = 10 million) and box-office collection, respectively, based on the data of 149 Bollywood movies (Data file: Bollywood Data.Xls).

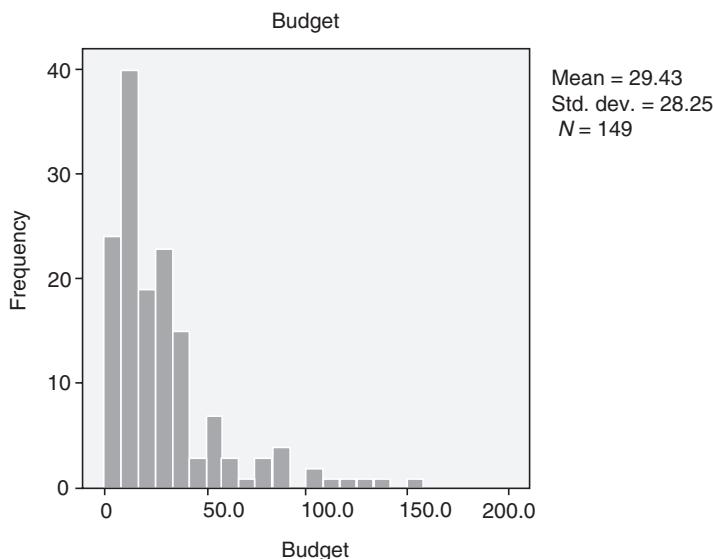
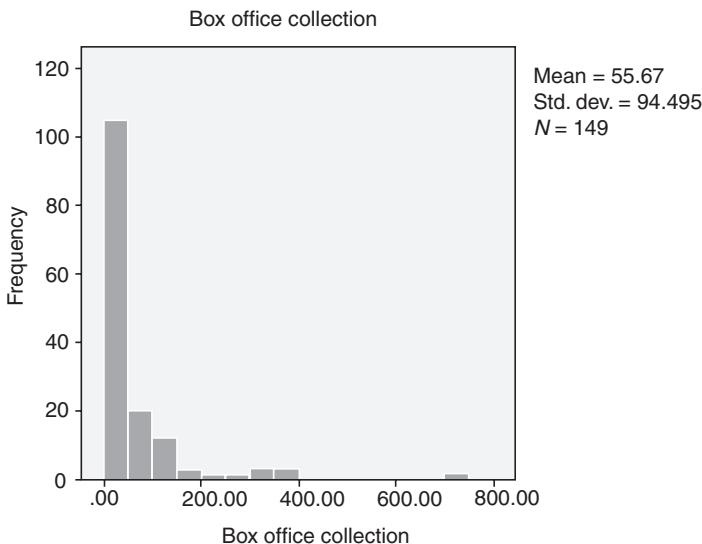


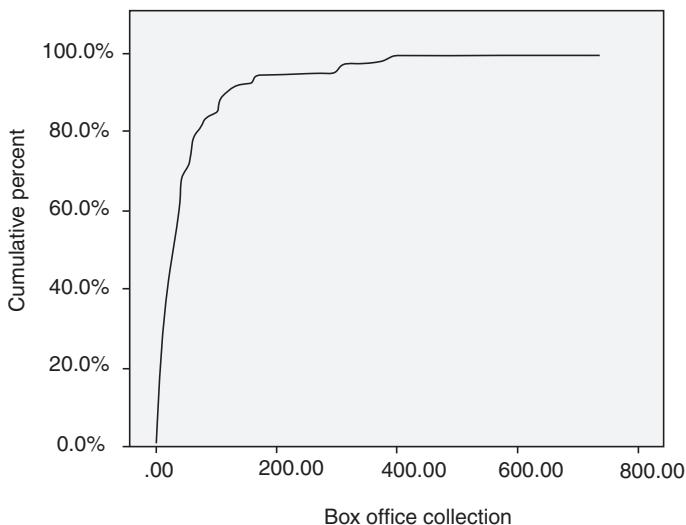
FIGURE 2.3 Histogram of Bollywood movie budget.



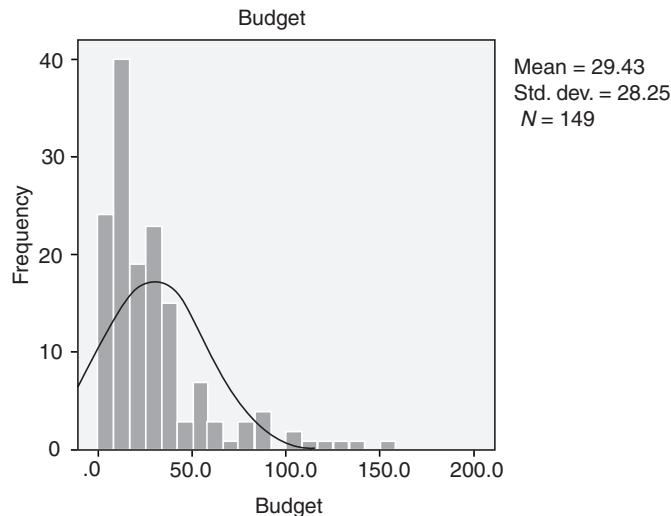
**FIGURE 2.4** Histogram of Bollywood movie box-office collection.

From Figure 2.3, we can infer that the budget for large proportion of movies is less than 50 crores and it is a right-skewed distribution (that is, long tail on the right-hand side). In Figure 2.4, we can also see an outlier where the box-office collection is more than 700 crores (movie PK acted by Amir Khan and directed by Rajkumar Hirani). The cumulative histograms are called **Ogive curves**. The Ogive curve for Bollywood box-office collection is shown in Figure 2.5.

Usually, we superimpose normal distribution on the histogram to see how close the frequency distribution of the data is to a normal distribution. Figure 2.6 shows histogram of movie budget superimposed with normal distribution; it is obvious that the frequency distribution of budget is not a normal distribution.



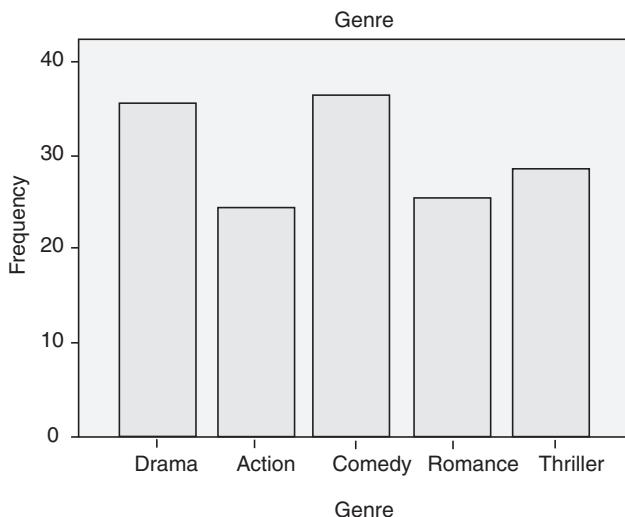
**FIGURE 2.5** Ogive curve for box-office collection.



**FIGURE 2.6** Histogram of Bollywood movie budget along with normal distribution frequency.

### 2.9.2 | Bar Chart

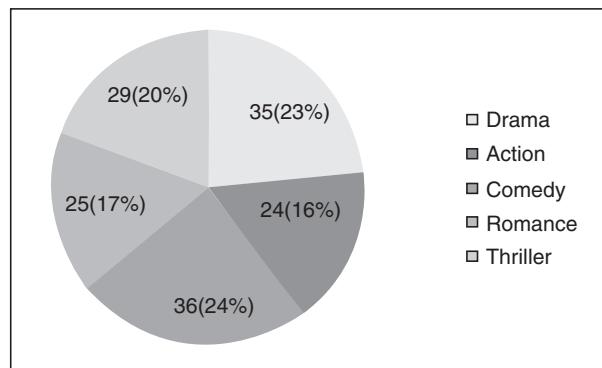
Bar chart is a frequency chart for qualitative variable (or categorical variable). Histograms cannot be used when the variable is qualitative. Bar chart can be used to assess the most-occurring and least-occurring categories within a data set. Figure 2.7 shows the bar chart for the movie genre (Data file: Bollywood Data.xlsx). From the bar chart, it is evident that genres, drama and comedy, are mostly preferred by the production houses in Bollywood.



**FIGURE 2.7** Bar chart for movie genre.

### 2.9.3 | Pie Chart

Pie chart is mainly used for categorical data and is a circular chart that displays the proportion of each category in the data set. The pie chart for the movie genre based on the Bollywood movie data set is shown in Figure 2.8.

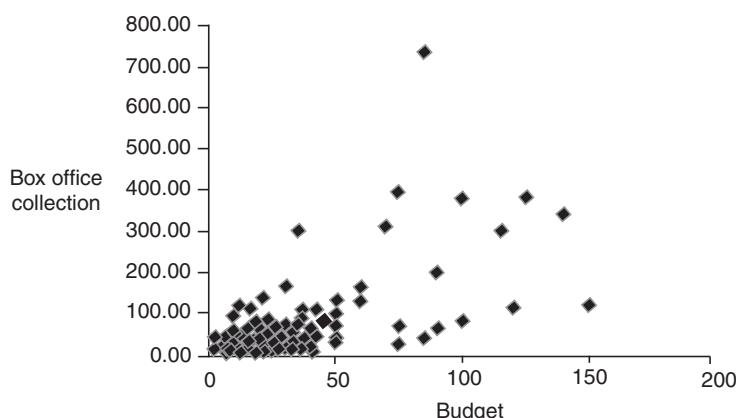


**FIGURE 2.8** Pie chart for movie genre.

Pie chart helps to visualize the proportion (percentage) of each category as sector of a circle.

### 2.9.4 | Scatter Plot

Scatter plot is a plot of two variables that will assist data scientists to understand if there is any relationship between two variables. The relationship could be linear or non-linear. Scatter plot is also useful for assessing the strength of the relationship and to find if there are any outliers in the data. Figure 2.9 shows a scatter plot between the movie budget and movie box-office collection (in crores of rupees) plotted using the data set in file Bollywood Data.xlsx.



**FIGURE 2.9** Scatter plot between movie budget and box-office collection.

Figure 2.9 shows a linear relationship between budget and box-office collection and existence of an outlier. Scatter plots are useful during regression model building to decide on the initial model, that is whether to consider a variable in a regression model or not.

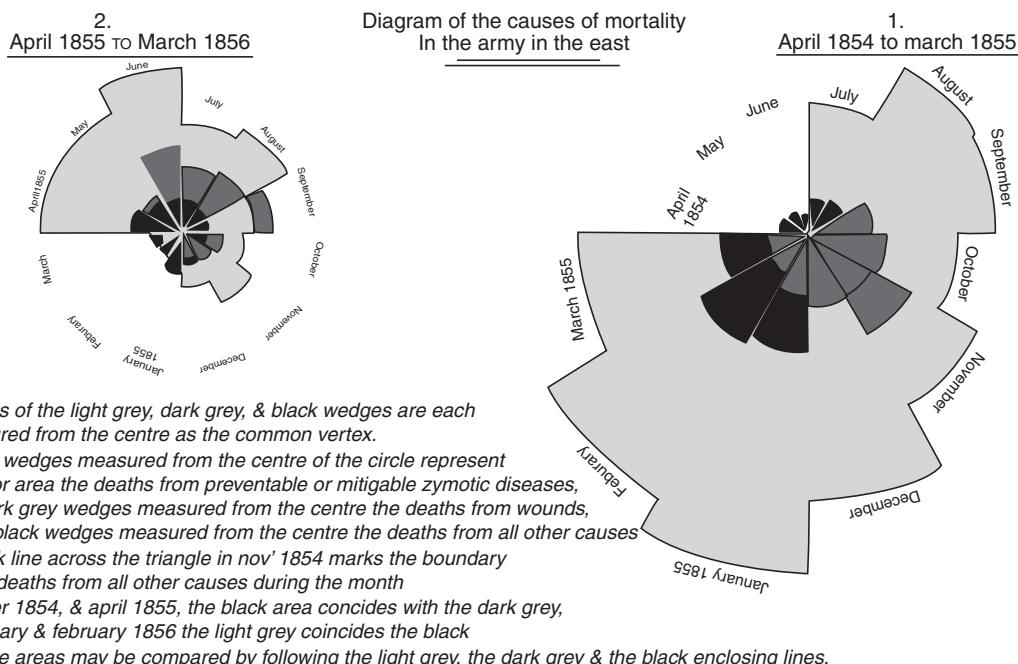
## 2.9.5 | Coxcomb Chart

Coxcomb chart (also known as polar area chart or roses) is an extension of pie chart made popular by Florence Nightingale (Lewi, 2006). In a Coxcomb chart, each area represents the magnitude of the category. The main difference between the regular pie chart and coxcomb chart is that in the case of pie chart the radius of each sector is same, whereas, in coxcomb chart the radius of the sector is adjusted to create the magnitude of the area.

Florence Nightingale collected data from Crimean war (war between British and French on one side and Russians on the other side) on causes of mortality among soldiers. She classified the causes into three categories:

1. Preventable diseases
2. Wounds sustained in the war
3. Other causes

In Figure 2.10 (originally prepared by Florence Nightingale), the largest area of the chart corresponds to the cause 'preventable diseases'.



**FIGURE 2.10** Coxcomb chart on causes of mortality in the army prepared by Florence Nightingale.<sup>1</sup>

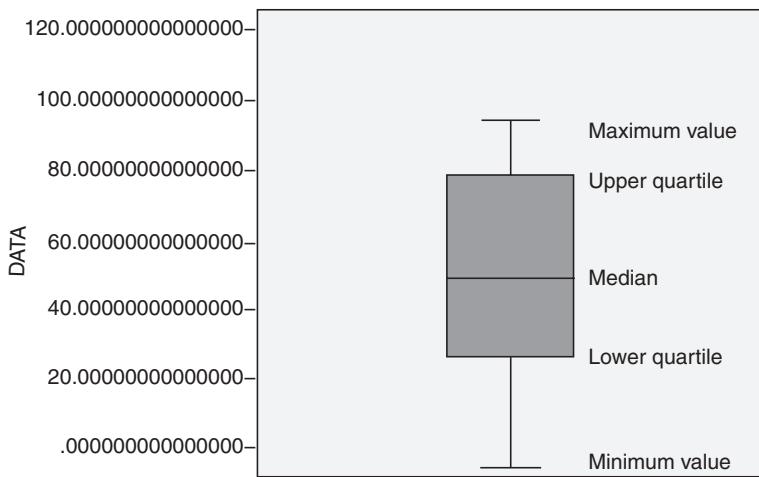
<sup>1</sup> Source: [https://en.wikipedia.org/wiki/Florence\\_Nightingale#/media/File:Nightingale-mortality.jpg](https://en.wikipedia.org/wiki/Florence_Nightingale#/media/File:Nightingale-mortality.jpg)

## 2.9.6 | Box Plot (or Box and Whisker Plot)

Box plot (aka Box and Whisker plot) is a graphical representation of numerical data that can be used to understand the variability of the data and the existence of outliers. Box plot is designed by identifying the following descriptive statistics:

1. Lower quartile (1<sup>st</sup> Quartile), median and upper quartile (3<sup>rd</sup> Quartile).
2. Lowest and highest value.
3. Inter-quartile range (IQR).

The box plot is constructed using IQR, minimum and maximum values. The box plot for the data in Table 2.4 is shown in Figure 2.11.



**FIGURE 2.11** Box plot of the data in Table 12.3.

The length of the box is equivalent to IQR. It is possible that the data may contain values beyond  $Q_1 - 1.5 \text{ IQR}$  and  $Q_3 + 1.5 \text{ IQR}$ . The whisker of the box plot extends till  $Q_1 - 1.5 \text{ IQR}$  (or minimum value) and  $Q_3 + 1.5 \text{ IQR}$  (or maximum value); observations beyond these two limits are potential outliers. The box plot for the Bollywood movie budget is shown in Figure 2.12.

In Figure 2.12 position of the lowest whisker is 2 (since that is the minimum value). The value of lower quartile is 11 (lower line of the box), median is 24 (middle line in the box), and top quartile is 35 (upper line of the box). The top whisker is at  $Q_3 + 1.5 \text{ IQR} = 71$ . All the observations beyond  $Q_3 + 1.5 \text{ IQR}$  shown above the upper whisker are outliers.

## 2.9.7 | Treemap

Treemap is a hierarchical map made up of nested rectangles frequently used as part of business intelligence reports which helps organizations to understand the data hierarchically. To construct a treemap, the data should be hierarchical with several levels. The size of rectangle and colour are used for describing/differentiating the characteristics of the data. A sample Treemap is shown in Figure 2.13 in which

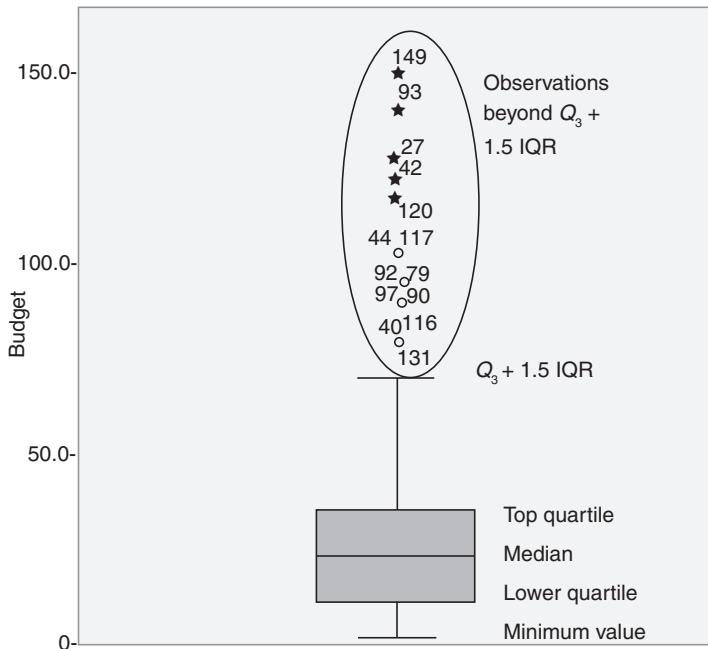


FIGURE 2.12 Box plot for Bollywood movie budget.

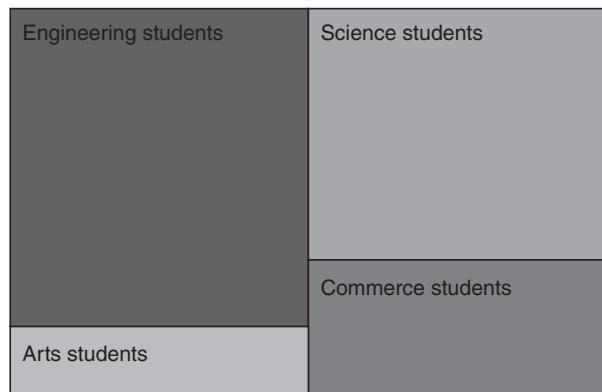


FIGURE 2.13 Treemap of student discipline at the undergraduate level.

the academic discipline at undergraduate level of students admitted into an MBA program is captured. Size of the area captures proportion of the students from that discipline. That is, in Figure 2.13 the area corresponding to engineering students is the largest indicating that the largest proportion of students come from engineering background and area corresponding to Arts students is the least indicating least number of students with arts background in the MBA program.

Each of the disciplines can be further analysed. For example, the engineering students can be further grouped according to the type of college (Tier 1, Tier 2, etc.).

**SUMMARY**

1. Descriptive analytics is beginning of any analytics project that uses data summarization, descriptive statistics, visualization and queries to gain insights about what happened in the past.
2. Measures of central tendency, measures of variation and measures of shape assist data scientists to understand the data for characteristics such as variability and skewness.
3. Data visualization is an integral part of descriptive analytics and plays a major role in business intelligence (BI) by displaying data using innovative graphs and dashboards for easy comprehension of data to top management.
4. Descriptive analytics can help data scientists with further analysis of the data by identifying relationships that may exist in the data.
5. Descriptive analytics will provide hints for developing predictive analytics models.

**MULTIPLE CHOICE QUESTIONS**

1. Which of the following variable is an interval scale variable?
  - (a) Age
  - (b) Latitude and longitude
  - (c) Marital status
  - (d) Hair colour
2. Which of the following measures are inappropriate for nominal scale variables?
  - (a) Mean
  - (b) Standard deviation
  - (c) Median
  - (d) Mode
3. Which of the following descriptive statistics is not calculated based on the entire data?
  - (a) Mean
  - (b) Skewness
  - (c) Kurtosis
  - (d) Median
4. Skewness is a measure of
  - (a) Location
  - (b) Scale
  - (c) Shape
  - (d) Range
5. The value of  $\sum_{i=1}^n (X_i - \bar{X})$  is
  - (a) Zero for any sample
  - (b) Zero for population, not necessarily for samples
  - (c) Zero for both samples and population
  - (d) Cannot say
6. For a positively skewed distribution
  - (a) Median is greater than mean
  - (b) Mean is greater than median
  - (c) Mode is greater than median and median is greater than mean
  - (d) Mean is greater than median and median is greater than mode
7. Ogive curves are
  - (a) Frequency charts
  - (b) Cumulative frequency charts
  - (c) Bar charts
  - (d) Cumulative bar charts
8. Bar charts are most relevant for
  - (a) Quantitative variable
  - (b) Qualitative variable
  - (c) Ordinal variable
  - (d) Interval scale variable
9. In a box plot the length of the box is
  - (a) Equal to variance of the data
  - (b) Equal to the standard deviation
  - (c) Equal to inter-quartile range (IQR)
  - (d) Equal to median + 1.5 IQR
10. In a box plot, an observation beyond  $Q_3 + 1.5 \text{ IQR}$  is
  - (a) A potential outlier
  - (b) Maximum value
  - (c) Mode
  - (d) Median
11. In a data set with 50 observations, two parameters were estimated. The corresponding degrees of freedom is
  - (a) 50
  - (b) 52
  - (c) 48
  - (d) between 48 and 52

12. In a sample variance, the sum of squared deviation from mean,  $\sum_{i=1}^n (X_i - \bar{X})^2$ , is divided by  $(n - 1)$  to overcome
- Upward bias of sum of squared deviation
  - Downward bias of sum of squared deviation
  - Lack of complete data
  - Bias in estimation of mean  $\bar{X}$

### EXERCISES

1. The daily footfall at a retail store in Bangalore over the last 30 days is shown in Table 2.7. Calculate the mean, median, mode and standard deviation.

**TABLE 2.7** Footfall data

232	277	261	173	283	197	251	212	213	213
229	164	219	196	186	247	244	269	216	272
252	314	161	165	221	260	219	290	225	251

2. For the data in Table 2.7, calculate the skewness and kurtosis. What can you infer from the skewness and kurtosis of the footfall data?
3. For the data in Table 2.7, calculate the values of first quartile and third quartile. Are there any outliers in the data?
4. The Bank of Kala Bakra (BKB) situated in Bakrapur, India receives several applications for home loan and home improvement loan. The description of the data captured in 'know your customer' (KYC) document is listed below (Data file: BKB.xls):
- Customer ID
  - Type of loan (2 types: Home Loan and Home Improvement Loan)
  - Gender (Male, Female)
  - Marital Status (Married, Single and Others)
  - Accommodation type (Family Other, Company Provided, Owned, Rented)
  - Number of years in the current address
  - Number of years in the current job
  - Monthly salary in Indian rupees
  - Balance in savings account (in Indian Rupees)
  - Loan amount requested (in Indian Rupees)
  - Term (loan term in months)
  - Down payment (in Indian rupees)
  - Equal Monthly Installment (EMI) affordable
- Develop appropriate charts for the variables. What insights can be obtained based on the charts?
  - Calculate the mean, median, mode, variance, standard deviation, skewness and kurtosis of variables monthly salary and balance in saving account.
  - Use box plot to check whether there are outliers among variables loan amount requested, down payment, and EMI.
  - Which variable among continuous variables have high skewness?
5. The cumulative grade point average (CGPA) of 40 students are shown in Table 2.8.

**TABLE 2.8** CGPA of students

3.36	1.56	1.48	1.43	2.64	1.48	2.77	2.20	1.38	2.84
1.88	1.83	1.87	1.95	3.43	1.28	3.67	2.23	1.71	1.68
2.57	3.74	1.98	1.66	1.66	2.96	1.77	1.62	2.74	3.35
1.80	2.86	3.28	1.14	1.98	2.96	3.75	1.89	2.16	2.07

- (a) Calculate the mean, median and mode. Calculate the standard deviation.
- (b) Calculate the 90<sup>th</sup> and 95<sup>th</sup> percentile of CGPA.
- (c) Calculate the inter quartile range (IQR).
- (d) The Dean of the school believes that the CGPA is a right tailed distribution. Is there an evidence to support dean's belief?
- (e) Create a histogram for the data, what should be the ideal number of bins in the histogram.
6. Value of insurance claims at an insurance company has mean value of INR 7200 and standard deviation of 200. Comment on the proportion of claims with values between INR 6900 and 7500.
7. Share Khan and Sons (SKS) is an investment advisory company. SKS has identified top 50 shares and its value rounded to nearest rupees are shown in Table 2.9.

**TABLE 2.9** Value of shares in rupees

600	349	292	247	216	411	233	364	419	505
474	541	790	293	362	470	349	429	565	309
453	419	354	273	533	235	467	569	590	347
413	541	318	545	256	247	474	597	522	535
483	573	345	568	260	288	50	248	466	417

- (a) Plot a histogram for the data. What insights can you gain from the histogram?
- (b) Plot a box plot and identify if there are any outliers.
- (c) Is the distribution of share price mesokurtic? Respond using an appropriate measure.
- (d) If the value of share is 600, calculate its percentile value.
8. Demand for a spare parts sold by a capital equipment manufacturer is shown in Table 2.10.

**TABLE 2.10** Demand for spare parts

2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
65	106	55	98	80	84	105	111	103	137

- (a) What type of data is provided in Table 2.10 (cross-sectional, time series, or panel)?
- (b) Use an appropriate chart for the data in Table 2.10. What insights you can get from the used chart?

## REFERENCES

1. Joanes D N and Gill C A (1998), "Comparing Measures of Sample Skewness and Kurtosis", *Journal of Royal Statistical Society – Series D*, 47(1), 183–189.
2. Lewi P (2006), "Speaking of Graphics", available at <http://www.datascope.be/sog/SOG-Title.pdf> accessed on 10 April 2017.

3. Sturges S A (1926), “The Choice of Class Interval”, *Journal of the American Statistical Association*, **21**, 65–66, 1926.
4. Surowiecki J (2004), “*The Wisdom of Crowds*”, Doubleday, New York.
5. Pandey S and Bright C L (2008), “What are Degrees of Freedom?”, *Social Work Research*, **32**(2), 119–128 (2008)
6. Toothaker L E and Miller L (1996), “*Introductory Statistics for Behavioural Sciences* (2<sup>nd</sup> Edition)”, Pacific Grove, California.
7. Trochim W M K (2005), “*Research Methods: Concise Knowledge Base*”, Atomic Dog Publishing Inc, Mason, Ohio.
8. Walker H W (1940), “Degrees of Freedom”, *Journal of Educational Psychology*, **31**, 253–269.

# Introduction to Probability

3

“To understand God’s thoughts we must study Statistics,  
for these are measure of His purpose.”

— Florence Nightingale

## LEARNING OBJECTIVES

- LO 3-1** Understand uncertainty and how probability concepts are used for measuring and modelling uncertainty.
- LO 3-2** Learn basic concepts in probability: axioms of probability, frequency estimate of probability, conditional probability and Bayes’ theorem.
- LO 3-3** Learn how simple probability rules are used for solving business problems using association rule mining and its applications in market basket analysis and recommender systems.
- LO 3-4** Understand the concept of random variables, discrete and continuous random variables, probability density function, and cumulative distribution function.
- LO 3-5** Understand central limit theorem and its importance in analytics.
- LO 3-6** Understand various discrete distributions such as binomial distribution, Poisson distribution, and geometric distribution and their applications for solving business problems.
- LO 3-7** Understand various continuous distributions such as uniform, exponential, normal, chi-square, *t*, and *F* distributions and their applications for solving business problems.

## IMPORTANCE OF PROBABILITY THEORY IN ANALYTICS

One of the primary objectives in analytics is to measure the uncertainty associated with an event or key performance indicator. Axioms of probability and the concept of random variable are fundamental building blocks of analytics that are used for measuring uncertainty associated with key performance indicators of importance for a business. Probability theory is the foundation on which descriptive and predictive analytics models are built.

IMPORTANT

*Understanding of probability concepts is important for analytics model building.*

### 3.1 | INTRODUCTION TO PROBABILITY THEORY

Analytics applications involve tasks such as prediction of probability of occurrence of an event, testing a hypothesis, building models to explain variation in a variable of importance to the business such as

profitability, market share, demand, etc. Many important tasks in analytics deal with uncertain events and it is essential to understand probability theory that can be used to predict and measure uncertain events. In this chapter, we will be discussing axioms of probability, concept of random variables, discrete and continuous probability distributions, and how each of these concepts are used in solving different analytics problems. This chapter is not intended as a rigorous treatment of all-relevant theorems and proofs. The intention is to provide an understanding of the main concepts in probability theory that forms the basis for predictive analytics.

## 3.2 | PROBABILITY THEORY – TERMINOLOGY

In this section, we will be discussing various terminologies that are used in probability theory.

### 3.2.1 | Random Experiment

Random experiment is an experiment in which the outcome is not known with certainty. That is, the output of a random experiment cannot be predicted with certainty. Predictive analytics mainly deals with random experiments such as predicting quarterly revenue of an organization, customer churn (whether a customer is likely to churn or how many customers are likely to churn before next quarter), demand for a product at a future time period, number of views for an YouTube video, outcome of a football match (win, draw or lose), etc.

### 3.2.2 | Sample Space

Sample space is the universal set that consists of all possible outcomes of an experiment. Sample space is usually represented using the letter 'S' and individual outcomes are called the elementary events. The sample space can be finite or infinite. Few random experiments and their sample spaces are discussed below:

**Experiment:** Outcome of a football match

$$\text{Sample Space} = S = \{\text{Win, Draw, Lose}\}$$

**Experiment:** Predicting customer churn at an individual customer level

$$\text{Sample Space} = S = \{\text{Churn, No Churn}\}$$

**Experiment:** Predicting percentage of customer churn

$\text{Sample Space} = S = \{X \mid X \in R, 0 \leq X \leq 100\}$ , that is  $X$  is a real number that can take any value between 0 and 100 percentage.

**Experiment:** Life of a turbine blade used in an aircraft engine

$\text{Sample Space} = S = \{X \mid X \in R, 0 \leq X < \infty\}$ , that is  $X$  is a real number that can take any value between 0 and  $\infty$ .

### 3.2.3 | Event

Event ( $E$ ) is a subset of a sample space and probability is usually calculated with respect to an event. An event can be represented using the Venn diagram in Figure 3.1.

The Venn diagram in Figure 3.1 indicates that the event  $E$  is a subset of the sample space  $S$ , that is,  $E \subset S$  ( $E$  is a subset of  $S$ ). Consider the random experiment that predicts number of customers who are likely to churn within a quarter from a customer base of 100 customers. The corresponding sample space =  $\{X | X \in Z, 0 \leq X \leq 100\}$ , that is  $X$  is a real number that can take any integer value between 0 and 100. Now we can define several events such as:

Event  $A$  = Number of customer churn less than 10

Event  $B$  = Number of customer churn between 10 and 30

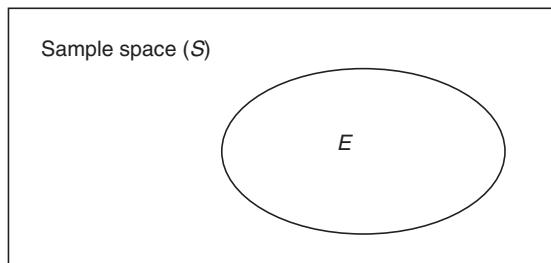
Event  $C$  = Number of customer churn exceeding 30

Consider another random experiment that predicts number of items purchased by a customer in a retail store and assume that the store has 5000 stock keeping units (SKUs). The sample space  $S = \{1, 2, \dots, 5000\}$ . This random experiment can have several events such as:

Event  $X$  = Number of SKUs purchased is less than 5

Event  $Y$  = Number of SKUs purchased is more than 20

Event  $Z$  = Number of SKUs purchased is between 10 and 20.



**FIGURE 3.1** Sample space ( $S$ ) and event ( $E$ ).

### 3.2.4 | Probability Estimation using Relative Frequency

The classical approach to probability estimation of an event is based on the relative frequency of the occurrence of that event. According to frequency estimation, the probability of an event  $X$ ,  $P(X)$ , is given by

$$P(X) = \frac{\text{Number of observations in favour of event } X}{\text{Total number of observations}} = \frac{n(X)}{N} \quad (3.1)$$

For example, say a company has 1000 employees and every year about 200 employees leave the job. Then the probability of attrition of an employee per annum is  $200/1000 = 0.2$ .

**EXAMPLE 3.1**

A website displays 10 advertisements and the revenue generated by the website depends on the number of visitors to the site clicking on any of the advertisements displayed on the website. The data collected by the company has revealed that out of 2500 visitors, 30 visitors clicked on 1 advertisement, 15 clicked on 2 advertisements, and 5 clicked on 3 advertisements. Remaining did not click on any of the advertisements. Calculate

- The probability that a visitor to the website will click on an advertisement.
- The probability that the visitor will click on at least two advertisements.
- The probability that a visitor will not click on any advertisements.

**Solution:**

- Number of customers clicking an advertisement is 50 and the total number of visitors is 2500. Thus, the probability that a visitor to the website will click on an advertisement is

$$\frac{50}{2500} = 0.02$$

- Number of customers clicking on at least 2 advertisements is 20. Thus, the probability that a visitor will click on at least 2 advertisements is

$$\frac{20}{2500} = 0.008$$

- Probability that a visitor will not click on any advertisement is

$$\frac{2450}{2500} = 0.98$$

### 3.2.5 | Algebra of Events

Assume that  $X$ ,  $Y$  and  $Z$  are three events of a sample space. Then the following algebraic relationships are valid and are useful while deriving probabilities of events:

**Commutative rule:**  $X \cup Y = Y \cup X$  and  $X \cap Y = Y \cap X$

**Associative rule:**  $(X \cup Y) \cup Z = X \cup (Y \cup Z)$  and  $(X \cap Y) \cap Z = X \cap (Y \cap Z)$

**Distributive rule:**  $X \cup (Y \cap Z) = (X \cup Y) \cap (X \cup Z)$

$X \cap (Y \cup Z) = (X \cap Y) \cup (X \cap Z)$

The above rules of algebra will be useful while calculating the probability of events. The following rules known as *DeMorgan's Laws* on complementary sets are useful while deriving probabilities:

$$(X \cup Y)^c = X^c \cap Y^c$$

$$(X \cap Y)^c = X^c \cup Y^c$$

where  $X^c$  and  $Y^c$  are the complementary events of  $X$  and  $Y$ , respectively.

### 3.3 | FUNDAMENTAL CONCEPTS IN PROBABILITY – AXIOMS OF PROBABILITY

In 1933, Andrey Kolmogorov, a Russian mathematician laid the foundation of the axiomatic theory of probability (Kolmogorov, 1956). According to axiomatic theory of probability, the probability of an event  $E$  satisfies the following axioms:

1. The probability of event  $E$  always lies between 0 and 1. That is,  $0 \leq P(E) \leq 1$ .
2. The probability of the universal set  $S$  is 1. That is,  $P(S) = 1$ .
3.  $P(X \cup Y) = P(X) + P(Y)$ , where  $X$  and  $Y$  are two mutually exclusive events.

Using the aforementioned axioms of probability, one can derive several mathematical relationships on probability of events using set theory logic. The following elementary rules of probability are directly deduced from the original three axioms of probability, using the set theory relationships:

1. For any event  $A$ , the probability of the complementary event, written  $A^C$ , is given by

$$P(A^C) = 1 - P(A) \quad (3.2)$$

If  $P(A)$  is a probability of observing a fraudulent transaction at an e-commerce portal, then  $P(A^C)$  is the probability of observing a genuine transaction.

2. The probability of an empty or impossible event,  $\emptyset$ , is zero:

$$P(\emptyset) = 0 \quad (3.3)$$

3. If occurrence of an event  $A$  implies that an event  $B$  also occurs, so that the event class  $A$  is a subset of event class  $B$ , then the probability of  $A$  is less than or equal to the probability of  $B$ :

$$P(A) \leq P(B) \quad (3.4)$$

If  $A$  is students with more than 3.5 CGPA (cumulative grade point average) out of 4 and  $B$  is students with a CGPA of more than 3.0, then  $P(A) \leq P(B)$ .

4. The probability that either events  $A$  or  $B$  occur or both occur is given by

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (3.5)$$

5. If  $A$  and  $B$  are mutually exclusive events, so that  $P(A \cap B) = 0$ , then

$$P(A \cup B) = P(A) + P(B) \quad (3.6)$$

6. If  $A_1, A_2, \dots, A_n$  are  $n$  events that form a partition of sample space  $S$ , then their probabilities must add up to 1:

$$P(A_1) + P(A_2) + \dots + P(A_n) = \sum_{i=1}^n P(A_i) = 1 \quad (3.7)$$

#### 3.3.1 | Joint Probability

Let  $A$  and  $B$  be two events in a sample space. Then the joint probability of the two events, written as  $P(A \cap B)$ , is given by

$$P(A \cap B) = \frac{\text{Number of observations in } A \cap B}{\text{Total number of observations}} \quad (3.8)$$

**EXAMPLE 3.2**

At an e-commerce customer service centre a total of 112 complaints were received. 78 customers complained about late delivery of the items and 40 complained about poor product quality.

- (a) Calculate the probability that a customer complaint will be about both late delivery and product quality.
- (b) What is the probability that a complaint is only about poor quality of the product?

**Solution:**

Let  $A$  = Late delivery and  $B$  = Poor quality of the product. Let  $n(A)$  and  $n(B)$  be the number of cases in favour of  $A$  and  $B$ . So  $n(A) = 78$  and  $n(B) = 40$ . Since the total number of complaints is 112 (here complaints is treated as the sample space), hence

$$n(A \cap B) = 118 - 112 = 6$$

Probability of a complaint about both delivery and poor product quality is

$$P(A \cap B) = \frac{n(A \cap B)}{\text{Total number of complaints}} = \frac{6}{112} = 0.0535$$

Probability that the complaint is only about poor quality =  $1 - P(A) = 1 - \frac{78}{112} = 0.3035$

**EXAMPLE 3.3**

Table 3.1 describes loan default status at a bank and their marital status. Calculate the marital status that has maximum joint probability of default.

**TABLE 3.1** Joint and marginal probability

Marital Status	Loan Status		Total
	Default	Non-Default	
Single	42	258	300
Married	60	590	650
Divorced	13	37	50
Total	115	885	1000

**Solution:**

$$P(\text{Single} \cap \text{Default}) = 0.042$$

$$P(\text{Married} \cap \text{Default}) = 0.06$$

$$P(\text{Divorced} \cap \text{Default}) = 0.013$$

The maximum joint probability is for  $P(\text{Married} \cap \text{Default})$ .

### 3.3.2 | Marginal Probability

Marginal probability is simply a probability of an event  $X$ , denoted by  $P(X)$ , without any conditions. In Example 3.3, let

$$X_1 = \text{Loan Status Default}$$

$$X_2 = \text{Loan Status Non-default}$$

$$Y_1 = \text{Marital Status Single}$$

$$Y_2 = \text{Marital Status Married}$$

$$Y_3 = \text{Marital Status Divorced}$$

Then marginal probabilities are

$$P(X_1) = \frac{115}{1000} = 0.115, P(X_2) = \frac{885}{1000} = 0.885$$

$$P(Y_1) = \frac{300}{1000} = 0.3, P(Y_2) = \frac{650}{1000} = 0.65, P(Y_3) = \frac{50}{1000} = 0.05$$

### 3.3.3 | Independent Events

Two events  $A$  and  $B$  are said to be independent when occurrence of one event (say event  $A$ ) does not affect the probability of occurrence of the other event (event  $B$ ). Mathematically, two events  $A$  and  $B$  are independent when  $P(A \cap B) = P(A) \times P(B)$ . For example, winning the toss by the Indian cricket team captain in consecutive matches are independent events. Whereas, let event  $A$  be life of an equipment exceeding 100 hours and event  $B$  be life of the equipment exceeding 200 hours. Then events  $A$  and  $B$  are dependent events. Independent events are useful property of events since it simplifies calculating probability values.

### 3.3.4 | Conditional Probability

If  $A$  and  $B$  are events in a sample space, then the conditional probability of the event  $B$  given that the event  $A$  has already occurred, denoted by  $P(B|A)$ , is defined as

$$P(B|A) = \frac{P(A \cap B)}{P(A)}, \quad P(A) > 0 \quad (3.9)$$

The conditional probability symbol  $P(B|A)$  is read as the probability of  $B$  given  $A$ . It is necessary to satisfy the condition that  $P(A) > 0$ , because it does not make sense to consider the probability of  $B$  given that event  $A$  is impossible.

Based on Table 3.1, the conditional probability of default given divorced is

$$P(\text{Default}|\text{Divorced}) = 0.013/0.05 = 0.26$$

and similarly probability of default given single is

$$P(\text{Default}|\text{Single}) = 0.042/0.3 = 0.14$$

### 3.4 | APPLICATION OF SIMPLE PROBABILITY RULES – ASSOCIATION RULE LEARNING

We can use simple probability concepts such as joint probability and conditional probability to solve analytics problems such as market basket analysis and recommender systems using algorithms such as Association Rule Learning (aka Association Rule Mining). Association rule mining is one of the popular algorithms used to solve problems such as *market basket analysis* and *recommender systems*.

Market basket analysis (MBA) is used frequently by retailers to predict products a customer is likely to buy together, which further can be used for designing planogram and product promotions. The primary objective of MBA is to find probability of buying two products ( $A$  and  $B$ ) together. That is, if a customer buys a product (say chips), then the customer is also likely to buy soft drink (such as Coca Cola or Pepsi). The legendary example that many data mining professionals discuss is ‘beer and diaper story’, that is, customers who purchased beer also purchased baby diapers (Rao, 1998). Recommender systems are models that produce list of recommendations to a customer on products such as books, movies, news items, etc. and is an important analytics technique. As discussed in Chapter 1, companies such as Amazon and Netflix benefitted significantly by using recommender systems.

#### 3.4.1 | Association Rule Learning

In general, association rule learning (also known as association rule mining) is a method of finding association between different entities in a database. In a retail context, association rule learning is a method for finding association relationships that exist in frequently purchased items. Association rule is a relationship of the form  $X \rightarrow Y$  (that is,  $X$  implies  $Y$ ). Here,  $X$  and  $Y$  are two mutually exclusive sets (set of stock keeping units or SKUs). Association rules can be created using the point of sale (PoS) data from retail stores. Before generating the association rules, the data may be pre-processed and a new table is created by using binary code as shown in Table 3.2.

**TABLE 3.2** Binary representation of point of sale data

Transaction ID	Apple	Orange	Grapes	Strawberry	Plums	Green Apple	Banana
1	1	1	1	0	1	1	1
2	0	1	0	0	0	1	1
3	0	0	0	0	0	1	1
4	1	0	0	0	1	0	0
5	1	0	0	0	1	1	1
6	0	1	1	0	0	0	1
7	0	1	1	0	0	0	1

In Table 3.2, transaction ID is the transaction reference number and apple, orange, etc. are the different SKUs sold by the store. Binary code is used to represent whether the SKU was purchased (equal to 1) or not (equal to 0) during a transaction. The strength of association between two mutually exclusive subsets can be measured using ‘support’, ‘confidence’, and ‘lift’.

Support between two sets (of products purchased) is calculated using the joint probability of those events:

$$\text{Support} = P(X \cap Y) = \frac{n(X \cap Y)}{N} \quad (3.10)$$

where  $n(X \cap Y)$  is the number of times both  $X$  and  $Y$  is purchased together and  $N$  is the total number of transactions. That is, support is proportion of times  $X$  and  $Y$  are purchased together.

Confidence is the conditional probability of purchasing product  $Y$  given the product  $X$  is purchased. It measures probability of event  $Y$  (customer buying a product  $Y$ ) given the event  $X$  has occurred (the customer has already purchased product  $X$ ). That is,

$$\text{Confidence} = P(Y | X) = \frac{P(X \cap Y)}{P(X)} \quad (3.11)$$

The third measure in association rule mining is lift, which is given by

$$\text{Lift} = \frac{P(X \cap Y)}{P(X) \times P(Y)} \quad (3.12)$$

Lift overcomes one of the disadvantages of using confidence. For example,  $P(X)$  could be very small, making it less attractive for MBA and recommendation among millions of SKUs that a retailer may be selling.

In Table 3.2, assume that  $X$  = Apple and  $Y$  = Banana. Then

$$\text{Support} = P(X \cap Y) = 2/7 = 0.285$$

$$\text{Confidence} = P(X \cap Y)/P(X) = 2/3 = 0.667$$

$$\text{Lift} = \frac{P(X \cap Y)}{P(X) \times P(Y)} = \frac{2/7}{(3/7) \times (6/7)} = 0.778$$

Association rules can be generated based on threshold values of support, confidence and lift. For example, assume that the cut-off for support is 0.25 and confidence is 0.5 (lift should be greater than 1). Then we can conclude that  $X$  implies  $Y$  (that is, purchase of apple implies purchase of banana, however this rule will be ineffective since lift is less than 1). Business rules can be generated such as what products to recommend to a customer based on support, confidence, and lift associated with the products already purchased by the customer and other products sold by the retailer.

Association rule mining is usually used on dyadic data that consists of two sets (say  $A$  and  $B$ ) and a pair [say  $(a, b)$ ] is formed such that  $a \in A$  and  $b \in B$ . For example, set  $A$  may represent set movies purchased by customers and set  $B$  is the feedback on those movies.

### 3.5 | BAYES' THEOREM

Bayes' theorem is one of the most important concepts in analytics since several problems are solved using Bayesian statistics. Consider two events  $A$  and  $B$ . We can write the following two conditional probabilities:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \text{ and } P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Using the two equations, we can show that

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \quad (3.13)$$

Equation (3.13) is the Bayes' theorem. Bayes' theorem helps the data scientists to update the probability of an event ( $B$ ) when any additional information is provided. This makes Bayesian statistics a very attractive technique since it helps the decision maker to fine-tune his/her belief with every additional data that is received. The following terminologies are used to describe various components in Eq. (3.13).

1.  $P(B)$  is called the *prior probability* (estimate of the probability without any additional information).
2.  $P(B|A)$  is called the *posterior probability* (that is, given that the event  $A$  has occurred, what is the probability of occurrence of event  $B$ ). That is, *post* the additional information (or additional evidence) that  $A$  has occurred, what is estimated probability of occurrence of  $B$ .
3.  $P(A|B)$  is called the likelihood of observing evidence  $A$  if  $B$  is true.
4.  $P(A)$  is the prior probability of  $A$ .

#### 3.5.1 | Solving Monty Hall Problem Using Bayes' Theorem

Let us revisit the Monty hall problem described in Chapter 1.

Let  $C_1$ ,  $C_2$ , and  $C_3$  be the events that the car is behind door 1, 2, and 3, respectively. Let  $D_1$ ,  $D_2$ , and  $D_3$  be the events that Monty opens door 1, 2, and 3, respectively. *Prior probabilities* of  $C_1$ ,  $C_2$ , and  $C_3$  are

$$P(C_1) = P(C_2) = P(C_3) = 1/3$$

Assume that the player has chosen door 1 and Monty opens door 2 to reveal a goat. Now we would like to calculate the posterior probability  $P(C_1|D_2)$ , that is, the probability that the car is behind door 1 (door chosen initially by the player) when Monty has provided the additional information that the car is not behind door 2. Using, Bayes' theorem

$$P(C_1|D_2) = \frac{P(D_2|C_1) \times P(C_1)}{P(D_2)} = \frac{(1/2) \times (1/3)}{(1/2)} = 1/3$$

$P(D_2|C_1) = \frac{1}{2}$  (if the car is behind door 1, then Monty can open either door 2 or 3)

$$P(D_2) = \frac{1}{2}$$

Note that  $P(C_2|D_2) = 0$ . Thus

$$P(C_3|D_2) = 1 - P(C_1|D_2) = 1 - \frac{1}{3} = \frac{2}{3}$$

Thus, changing the initial choice will increase the probability of winning the car. Alternatively,

$$P(C_3|D_2) = \frac{P(D_2|C_3) \times P(C_3)}{P(D_2)} = \frac{1 \times (1/3)}{(1/2)} = 2/3$$

$P(D_2|C_3) = 1$  (if the car is behind door 3 and the player has chosen door 1, Monty has to open door 2 with probability 1)

### 3.5.2 | Generalization of Bayes' Theorem

In Eq. (3.13), the probability of evidence  $P(A)$  may come from mutually exclusive subsets (events) as described in Figure 3.2.

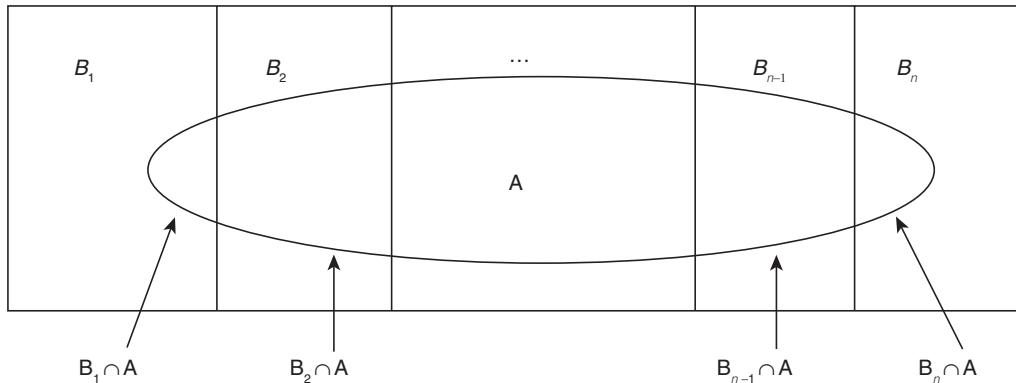


FIGURE 3.2 Event generated from mutually exclusive subsets.

For better understanding, consider a part manufactured by different suppliers  $B_1, B_2, \dots, B_n$ . Let  $A$  denote a defective part.  $P(A)$  can be written as

$$\begin{aligned} P(A) &= P(A, B_1) + P(A, B_2) + \dots + P(A, B_n) \\ &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \dots + P(A|B_n)P(B_n) \end{aligned} \quad (3.14)$$

where

$$P(A, B_1) = P(A \cap B_1).$$

**EXAMPLE 3.4**

Black boxes used in aircrafts are manufactured by three companies  $A$ ,  $B$  and  $C$ . 75% are manufactured by  $A$ , 15% by  $B$ , and 10% by  $C$ . The defect rates of black boxes manufactured by  $A$ ,  $B$ , and  $C$  are 4%, 6%, and 8%, respectively. If a black box tested randomly is found to be defective, what is the probability that it is manufactured by company  $A$ ?

**Solution:**

Let  $P(A)$ ,  $P(B)$ ,  $P(C)$  be events corresponding to the black box being manufactured by companies  $A$ ,  $B$ , and  $C$ , respectively, and  $P(D)$  be the probability of defective black box. We are interested in calculating the probability  $P(A|D)$ .

$$P(A|D) = \frac{P(D|A) \times P(A)}{P(D)}$$

Now  $P(D|A) = 0.04$  and  $P(A) = 0.75$ . Using Eq. (3.14):

$$P(D) = 0.75 \times 0.04 + 0.15 \times 0.06 + 0.10 \times 0.08 = 0.047$$

So

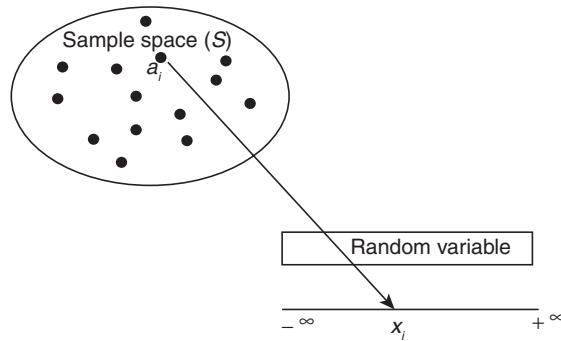
$$P(A|D) = \frac{0.04 \times 0.75}{0.047} = 0.6382$$

### 3.6 | RANDOM VARIABLES

Random variable is a function that maps every outcome in the sample space to a real number. Random variables provide robustness required while developing probabilistic models since the outcome of a random experiment may be recorded in different format. Outcomes of experiments may be recorded in numerical and non-numerical terms. For example, consider bank transactions that are classified as either genuine ( $G$ ) or fraud ( $F$ ). Assume that the bank looks at last four transactions; the sample space in this case can be written as  $S = \{GGGG, GGGF, GGFG, \dots\}$ . However, depending on the size of the bank, on any given day the number of transactions may run into several millions. Also, the bank would like to know the number of fraudulent transactions than the actual sequence of occurrence of fraudulent and genuine transactions. So, we need a variable that measures the number of fraudulent transactions. For every random experiment, we can define a function that maps the outcome to a real number. Random variable is defined as

A function that assigns a real number to each sample point in the sample space  $S$ .

The sample space discussed earlier  $S = \{GGGG, GGGF, GGFG, \dots\}$  will be mapped to a real number set  $S = \{0, 1, 2, 3, 4\}$ , which is often the interest of the analyst. In the set  $S = \{0, 1, 2, 3, 4\}$  the value represents the number of fraudulent transactions out of 4 transactions. Random variable is a robust and convenient way of representing the outcome of a random experiment. When the outcomes themselves are already expressed in terms of real numbers, it is possible to reassign a unique real number (or the original number itself) to the outcome. Random variables are usually denoted using capital letters, such as  $X$ ,  $Y$ , and  $Z$ , whereas small letters, such as  $x$ ,  $y$ ,  $z$ ,  $a$ ,  $b$ ,  $c$ , and so on, are used to denote particular values



**FIGURE 3.3** Random variable as a mapping from sample space to real number space.

of random variables written as  $P(X = x)$ . Figure 3.3 shows the mapping of the outcome of a random experiment to the real number. Once a random variable is defined, then the probability of events can be measured for various values that the random variable can take. For example, in the case of fraudulent transactions, we can now calculate probabilities such as

1.  $P(X = 2)$ , probability that the number of fraudulent transactions are exactly two.
2.  $P(X > 2)$ , probability that the number of fraudulent transactions are more than two.
3.  $P(X < 2)$ , probability that the number of fraudulent transactions are less than two.

Use of random variables provides us the flexibility required in modelling.

Random variables can be classified as discrete or continuous depending on the values that the random variable can take.

### 3.6.1 | Discrete Random Variables

If the random variable  $X$  can assume only a finite or countably infinite set of values, then it is called a discrete random variable. There are very many situations where the random variable  $X$  can assume only finite or countably infinite set of values. Examples of discrete random variables are:

1. Credit rating (usually classified into different categories such as low, medium and high or using labels such as AAA, AA, A, BBB, etc.).
2. Number of orders received at an e-commerce retailer which can be countably infinite.
3. Customer churn [the random variables take binary values: (a) Churn and (b) Do not churn].
4. Fraud [the random variables take binary values: (a) Fraudulent transaction and (b) Genuine transaction].
5. Any experiment that involves counting (for example, number of returns in a day from customers of e-commerce portals such as Amazon, Flipkart; number of customers not accepting job offers from an organization).

In analytics, classification problems, an important class of problems, is an example of discrete random variable.

### 3.6.2 | Continuous Random Variables

A random variable  $X$  which can take a value from an infinite set of values is called a continuous random variable. Examples of continuous random variables are listed below:

1. Market share of a company (which take any value from an infinite set of values between 0 and 100%).
2. Percentage of attrition among employees of an organization.
3. Time to failure of engineering systems.
4. Time taken to complete an order placed at an e-commerce portal.
5. Time taken to resolve a customer complaint at call and service centers.

In many situations, a continuous variable may be converted to a discrete random variable for modelling purpose.

### 3.6.3 | Probability Mass Function and Cumulative Distribution Function of a Discrete Random Variable

For a discrete random variable, the probability that a random variable  $X$  taking a specific value  $x_i$ ,  $P(X = x_i)$ , is called the probability mass function  $P(x_i)$ . That is, a probability mass function is a function that maps each outcome of a random experiment to a probability (Figure 3.4).

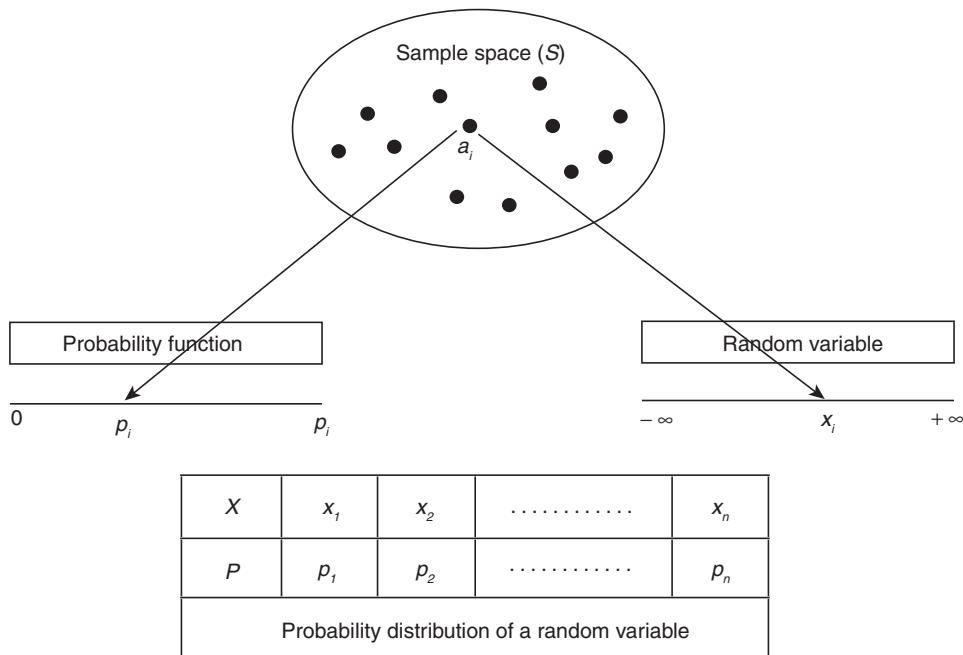


FIGURE 3.4 Probability mass function.

Consider the number of daily fraudulent transactions at a bank branch and the corresponding probabilities as described in Table 3.3. The values in Table 3.3 denote possible values of the random variable and the corresponding probability. That is, Table 3.3 describes how probability is distributed across different values of the random variable.

**TABLE 3.3** Probability mass function

Random Variable $X$ ( $X$ = number of fraudulent transactions)	$x_i = 0$	$x_i = 1$	$x_i = 2$	$x_i = 3$	$x_i = 4$
$P(X = x_i)$	0.20	0.15	0.25	0.25	0.15

From Table 3.3, we have the following information:

Probability that there will no fraudulent transaction on any given day,  $P(X = 0) = 0.20$ . Similarly,  $P(X = 1) = 0.15$ ,  $P(X = 2) = 0.25$ ,  $P(X = 3) = 0.25$  and  $P(X = 4) = 0.15$ .

The probability mass function,  $P(x_i)$  satisfies the following conditions:

1.  $P(x_i) \geq 0$ .
2.  $\sum_{x_i} P(x_i) = 1$

Cumulative distribution function,  $F(x_i)$ , is the probability that the random variable  $X$  takes values less than or equal  $x_i$ . That is,  $F(x_i) = P(X \leq x_i)$ .

Based on the values given in Table 3.3,

$$F(2) = P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = 0.60$$

### 3.6.4 | Expected Value, Variance, and Standard Deviation of a Discrete Random Variable

Expected value (or mean) of a discrete random variable is given by

$$E(X) = \sum_{i=1}^n x_i P(x_i) \quad (3.15)$$

where  $x_i$  is the specific value taken by a discrete random variable  $X$  and  $P(x_i)$  is the corresponding probability, that is,  $P(X = x_i)$ . Expected value of a discrete random variable plays a crucial role in many contexts. For example, expected monetary value (EMV) forms the basis for selecting an alternative from several possible alternatives in a decision tree approach. EMV is calculated based on expected value.

Variance of a discrete random variable is given by

$$\text{Var}(X) = \sum_{i=1}^n [x_i - E(X)]^2 \times P(x_i) \quad (3.16)$$

Standard deviation of a discrete random variable is given by

$$\sigma = \sqrt{\text{Var}(X)} \quad (3.17)$$

For the data in Table 3.2, the expected value of the number of fraudulent transactions is given by

$$E(X) = \sum_x x_i P(x_i) = 0 \times 0.2 + 1 \times 0.15 + 2 \times 0.25 + 3 \times 0.25 + 4 \times 0.15 = 2$$

That is, the average number of fraudulent transactions is 2.

The variance of the random variable  $X$  is given by

$$\begin{aligned} \text{Var}(X) &= \sum_{i=1}^n [x_i - E(X)]^2 \times P(x_i) \\ &= (0-2)^2 \times 0.2 + (1-2)^2 \times 0.15 + (2-2)^2 \times 0.25 + (3-2)^2 \times 0.25 + (4-2)^2 \times 0.15 \\ &= 1.8 \end{aligned}$$

The standard deviation,  $\sigma = \sqrt{\text{Var}(X)} = \sqrt{1.8} = 1.34$

### 3.7 | PROBABILITY DENSITY FUNCTION (PDF) AND CUMULATIVE DISTRIBUTION FUNCTION (CDF) OF A CONTINUOUS RANDOM VARIABLE

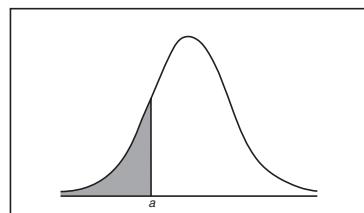
Since continuous random variables can take infinitely many values, their exact measurement is very difficult; even atomic clock comes with an error, although infinitesimally small. For this reason, the probability density function,  $f(x_i)$ , is defined as probability that the value of random variable  $X$  lies between an infinitesimally small interval defined by  $x_i$  and  $x_i + \delta x$  and its mathematical expression is

$$f(x) = \lim_{\delta x \rightarrow 0} \frac{P(x_i \leq X \leq x_i + \delta x)}{\delta x} \quad (3.18)$$

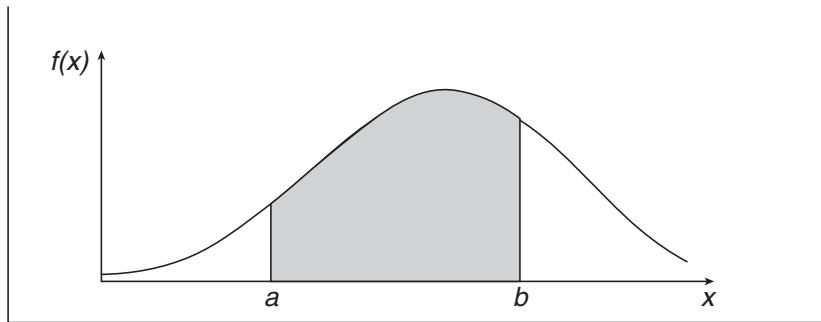
Probability density function reflects how dense is the likelihood of a continuous random variable  $X$  taking a value in an infinitesimally small interval around value  $x$ . The cumulative distribution function (CDF) of a continuous random variable is defined by

$$F(a) = P(X \leq a) = \int_{-\infty}^a f(x) dx \quad (3.19)$$

Cumulative distribution function  $F(a)$  is the area under the probability density function (Figure 3.5) up to  $X = a$ .



**FIGURE 3.5** Cumulative distribution function  $F(a)$ .



**FIGURE 3.6** Area between values  $(a, b)$  under probability density function.

Probability density function and cumulative distribution function of a continuous random variable satisfy the following properties:

1.  $f(x) \geq 0$

2.  $F(\infty) = \int_{-\infty}^{+\infty} f(x)dx = 1$

3.  $P(a \leq X \leq b) = \int_a^b f(x)dx = F(b) - F(a)$

The probability between two values  $a$  and  $b$ ,  $P(a \leq X \leq b)$ , is the area between the values  $a$  and  $b$  under the probability density function (Figure 3.6).

The expected value of a continuous random variable,  $E(X)$ , is given by

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx \quad (3.20)$$

The variance of a continuous random variable,  $\text{Var}(X)$ , is given by

$$\text{Var}(X) = \int_{-\infty}^{\infty} [x - E(X)]^2 f(x)dx \quad (3.21)$$

### 3.8 | BINOMIAL DISTRIBUTION

Binomial distribution is one of the most important discrete probability distribution due to its applications in several contexts. A random variable  $X$  is said to follow a Binomial distribution when

1. The random variable can have only two outcomes *success* and *failure* (also known as Bernoulli trials).
2. The objective is to find the probability of getting  $k$  successes out of  $n$  trials.
3. The probability of success is  $p$  and thus the probability of failure is  $(1 - p)$ .
4. The probability  $p$  is constant and does not change between trials.

Success and failure are generic terminologies used in binomial distribution; based on the context, the interpretation will change (winning a lottery can be considered as success and not winning as failure). In analytics, the following are few example problems that can be associated with Binomial distribution:

1. Customer churn where the outcomes are: (a) Customer churn and (b) No customer churn.
2. Fraudulent insurance claims where the outcomes are: (a) Fraudulent claim and (b) Genuine claim.
3. Loan repayment default by a customer where the outcomes are: (a) Default and (b) No default.
4. Cart abandonment in e-commerce (a situation where the customer adds items to his/her cart but does not make the purchase), where the outcomes are: (a) Cart abandonment and (b) No cart abandonment.
5. Employee attrition at a company where the outcomes are: (a) The employee leaves (exits) the company and (b) The employee does not leave the company.

Any business context in which there are only two outcomes can be analysed using Binomial distribution.

### 3.8.1 | Probability Mass Function (PMF) of Binomial Distribution

The PMF of the Binomial distribution (probability that the number of success will be exactly  $x$  out of  $n$  trials) is given by

$$PMF(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad 0 \leq x \leq n \quad (3.22)$$

where

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \quad (3.23)$$

In Microsoft Excel, the function ‘BINOM.DIST( $x, n, p, \text{false}$ )’ can be used for calculating the probability mass function of a binomial distribution.

### 3.8.2 | Cumulative Distribution Function (CDF) of Binomial Distribution

CDF of a binomial distribution function,  $F(a)$ , representing the probability that the random variable  $X$  takes value less than or equal to  $a$ , is given by

$$F(a) = P(X \leq a) = \sum_{k=0}^a P(X = k) = \sum_{k=0}^a \binom{n}{k} p^k (1-p)^{n-k} \quad (3.24)$$

In Microsoft Excel, the function ‘BINOM.DIST( $x, n, p, \text{true}$ )’ can be used for calculating the cumulative distribution function of a binomial distribution. As an illustration of PMF and CDF of the binomial distribution are shown in Figures 3.7 and 3.8 for parameters  $n = 10$  and  $p = 0.5$ .

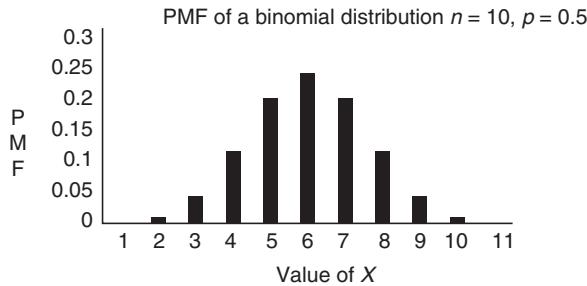


FIGURE 3.7 PMF of a binomial distribution.

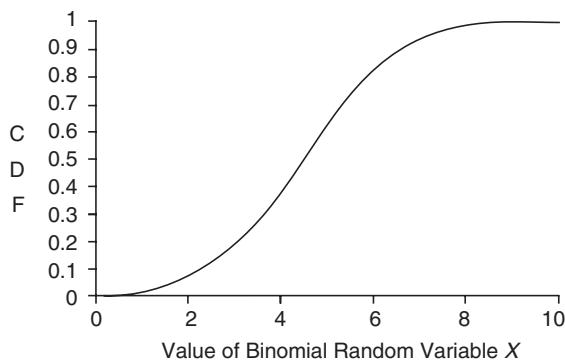


FIGURE 3.8 CDF of a binomial distribution.

### 3.8.3 | Mean and Variance of Binomial Distribution

Mean of a binomial distribution is given by

$$\text{Mean} = E(X) = \sum_{x=0}^n x \times \text{PMF}(x) = \sum_{x=0}^n x \times \binom{n}{x} p^x (1-p)^{n-x} = np \quad (3.25)$$

The variance of a binomial distribution is given by

$$\text{Var}(X) = \sum_{x=0}^n [x - E(X)]^2 \times \text{PMF}(x) = \sum_{x=0}^n [x - np]^2 \times \binom{n}{x} p^x (1-p)^{n-x} = np(1-p) \quad (3.26)$$

### 3.8.4 | Approximation of Binomial Distribution using Normal Distribution

If the number of trials ( $n$ ) in a binomial distribution is large, then it can be approximated by normal distribution with mean  $np$  and variance  $npq$ , where  $q = 1 - p$ .

**EXAMPLE 3.5**

Fashion Trends Online (FTO) is an e-commerce company that sells women apparel. It is observed that about 10% of their customers return the items purchased by them for many reasons (such as size, colour, and material mismatch). On a particular day, 20 customers purchased items from FTO. Calculate:

- Probability that exactly 5 customers will return the items.
- Probability that a maximum of 5 customers will return the items.
- Probability that more than 5 customers will return the items purchased by them.
- Average number of customers who are likely to return the items.
- The variance and the standard deviation of the number of returns.

**Solution:** In this case, the value of  $n = 20$  and  $p = 0.1$ .

- Probability that exactly 5 customers will return the items purchased is

$$P(X = 5) = \binom{20}{5} \times (0.1)^5 \times (0.9)^{15} = 0.03192$$

- Probability that a maximum of 5 customers will return the items purchased is

$$P(X \leq 5) = \sum_{k=0}^5 \binom{20}{k} \times (0.1)^k \times (0.9)^{20-k} = 0.9887$$

- Probability that more than 5 customers will return the product is

$$P(X > 5) = 1 - P(X \leq 5) = 1 - \sum_{k=0}^5 \binom{20}{k} \times (0.1)^k \times (0.9)^{20-k} = 1 - 0.9887 = 0.0113$$

- The average number of customers who are likely to return the items is

$$E(X) = n \times p = 20 \times 0.1 = 2$$

- Variance of a binomial distribution is given by

$$\text{Var}(X) = n \times p \times (1 - p) = 20 \times 0.1 \times 0.9 = 1.8$$

and the corresponding standard deviation is 1.3416.

**EXAMPLE 3.6**

Die Another Day (DAD) hospital recruits nurses frequently to manage high attrition among the nursing staff. Not all job offers from DAD hospital are accepted. Based on the past recruitment data, it was estimated that only 70% of offers rolled out by DAD hospital are accepted.

- If 10 offers are made, what is the probability that more than 5 and less than 8 candidates will accept the offer from DAD hospital?
- During March 2017, DAD required 14 new nurses to manage attrition. What should be the number of offers made by DAD hospital so that the average numbers of nurses accepting the offer is 14?

**Solution:**

- Probability that the number of accepted offers will be greater than 5 and less than 8 out of 10 offers is given by

$$P(5 < X < 8) = P(X = 6) + P(X = 7) = \binom{10}{6} \times 0.7^6 \times 0.3^4 + \binom{10}{7} \times 0.7^7 \times 0.3^3 = 0.4669$$

- For binomial distribution,  $E(X) = n \times p$ , and we have to find the number of offers such that on average 14 nurses accept the offer:

$$n \times p = 14 \Rightarrow n = \frac{14}{p} = \frac{14}{0.7} = 20$$

That is, the hospital should make 20 offers to ensure that the expected number of accepted offers is 14.

### 3.9 | POISSON DISTRIBUTION

In many situations, we may be interested in calculating the number of events that may occur over a period of time (or corresponding unit of measurement). For example, number of cancellation of orders by customers at an e-commerce portal, number of customer complaints, number of cash withdrawals at an ATM, number of typographical errors in a book, number of potholes on Bangalore roads, etc. When we have to find the probability of number of events, we use Poisson distribution. The probability mass function of a Poisson distribution is given by

$$P(X = k) = \frac{e^{-\lambda} \times \lambda^k}{k!}, \quad k = 0, 1, 2, \dots \quad (3.27)$$

where  $\lambda$  is the rate of occurrence of the events per unit of measurement (in many situations the unit of measurement is likely to be time). In Microsoft Excel, the function ‘POISSON.DIST( $k, \lambda, \text{false}$ )’ can be used for calculating the probability mass function of a Poisson distribution. Cumulative distribution function of a Poisson distribution is given by

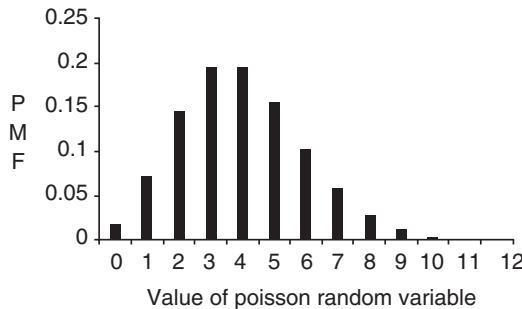


FIGURE 3.9 Probability mass function of a Poisson random variable ( $\lambda = 4$ ).

$$P[X \leq k] = \sum_{i=0}^k \frac{e^{-\lambda} \times \lambda^i}{i!} \quad (3.28)$$

In Microsoft Excel, the function ‘POISSON.DIST( $k, \lambda, \text{true}$ )’ can be used for calculating the cumulative distribution function of a Poisson distribution. The mean and variance of a Poisson random variable are given by

$$E(X) = \lambda \text{ and } \text{Var}(X) = \lambda \quad (3.29)$$

That is, mean and variance of a Poisson random variable are equal. Figures 3.9 and 3.10 show PMF and CDF of Poisson distribution for  $\lambda = 4$ .

If we are interested in predicting the number of events over a period of time, say between 0 and  $t$ , then the probability mass function is given by

$$P[N(t) = k] = \frac{e^{-\lambda t} \times (\lambda t)^k}{k!} \quad (3.30)$$

where  $N(t)$  is the number of events that occur over a period of time  $t$  or in  $[0, t]$ . The parameter  $\lambda$  is the mean number of occurrences per unit time and  $\lambda t$  is the mean number of occurrences in an interval  $[0, t]$ . We also assume that the events are independent.

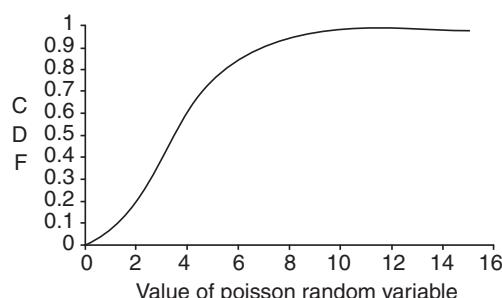


FIGURE 3.10 Cumulative distribution function of a Poisson random variable ( $\lambda = 4$ ).

**EXAMPLE 3.7**

On average, 20 customers per day cancel their order placed at Fashion Trends Online. Calculate the probability that the number of cancellations on a day is exactly 20 and the probability that the maximum number of cancellations is 25.

**Solution:**

The probability that the number of cancellations is exactly 20 is given by

$$P(X = 20) = \frac{e^{-20} 20^{20}}{20!} = 0.0888$$

Probability that the maximum number of cancellation will be 25 is given by

$$P(X \leq 25) = \sum_{k=0}^{25} \frac{e^{-20} 20^k}{k!} = 0.8878$$

**EXAMPLE 3.8**

The number of calls arriving at a call center follows a Poisson distribution at 10 per hour. Calculate the probability that the number of calls over a 3-hour period will exceed 30.

**Solution:**

Since the average calls per hour is 10 ( $\lambda = 10$ ), and we are interested in finding the calls over 3 hours, the mean number of calls over 3 hours is  $\lambda t = 30$ . Probability that the number of calls will be more than 30 is given by

$$P(X > 30) = 1 - P(X \leq 30) = 1 - \sum_{k=0}^{30} \frac{e^{-\lambda t} \times (\lambda t)^k}{k!} = 1 - \sum_{k=0}^{30} \frac{e^{-30} \times (30)^k}{k!} = 0.4516$$

### 3.10 | GEOMETRIC DISTRIBUTION

In many contexts, we may like to predict the number of failures before the occurrence of success in a Bernoulli trial. For example, in a repeat purchase such as grocery purchase, the grocery store may like to predict gaps between purchase of a specific product by the customer. This may be useful for developing recommender systems and ‘did you forget’ feature used by online grocers such as bigbasket.com discussed in Chapter 1. Geometric distribution represents a random experiment in which the random variable predicts the number of failures before the success. The probability density function of a geometric distribution is given by

$$P(X = x) = P(\text{success at } x^{\text{th}} \text{ trial}) = (1 - p)^{x-1} p, \quad \text{where } x = 1, 2, 3, \dots \quad (3.31)$$

where  $p$  is the probability of success. In Eq. (3.31) the success occurs at  $x^{\text{th}}$  trial. The cumulative distribution function is given by

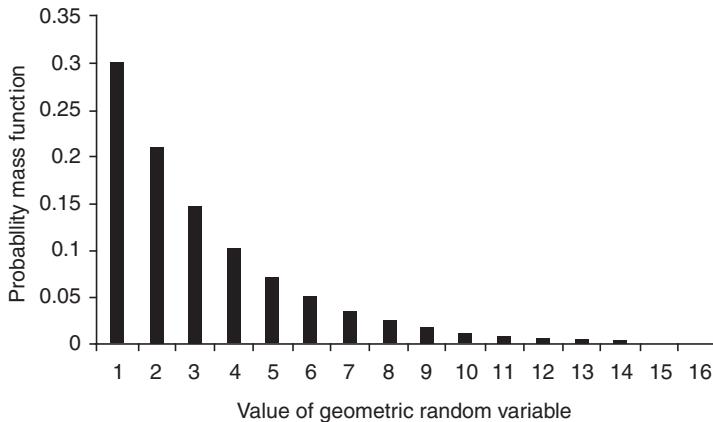


FIGURE 3.11 Probability mass function of a geometric distribution ( $p = 0.3$ ).

$$F(x) = P(X \leq x) = 1 - (1 - p)^x \quad (3.32)$$

Mean and variance of a geometric distribution are given by

$$E(X) = \frac{1}{p} \text{ and } \text{Var}(X) = \frac{(1-p)}{p^2} \quad (3.33)$$

The probability mass function and cumulative distribution function of a geometric distribution (for  $p = 0.3$ ) are shown in Figures 3.11 and 3.12.

### 3.10.1 | Memoryless Property of Geometric Distribution

Memoryless property is a special property of a geometric distribution in which the conditional probability,  $P(X > i + j | X > i)$ , depends only on the value  $j$ , not on the value  $i$ . We know that

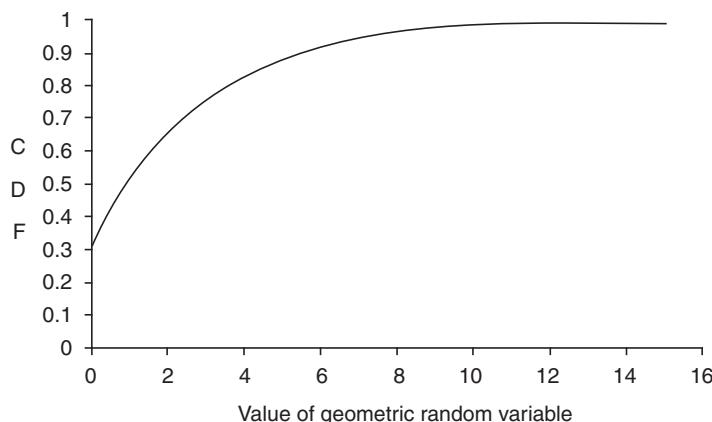


FIGURE 3.12 Cumulative distribution function of a geometric distribution ( $p = 0.3$ ).

$$P(X > i) = 1 - P(X \leq i) = 1 - [1 - (1-p)^i] = (1-p)^i$$

$$P(X > i+j | X > i) = \frac{P(X > i+j \cap X > i)}{P(X > i)} = \frac{P(X > i+j)}{P(X > i)} = \frac{(1-p)^{i+j}}{(1-p)^i} = (1-p)^j$$

Note that,

$$P(X > j) = (1-p)^j. \text{ Thus, } P(X > i+j | X > i) = P(X > j).$$

Memoryless property is an important property that simplifies calculations associated with conditional probabilities. Geometric distribution is the only discrete probability distribution that has the memoryless property.

### EXAMPLE 3.9

Local Dhaniawala (LD) is an online grocery store and has an innovative feature which predicts whether the customer has forgotten to buy an item which is very common among customers of grocery items. The probability that a customer buys milk in each shopping visit is 0.2.

- (a) Calculate the probability that the customer's first purchase of milk happens during the 5<sup>th</sup> visit.
- (b) Calculate the average time between purchases of milk.
- (c) If a customer has not purchased milk during the past 3 shopping visits, what is the probability that the customer will not buy milk for another 2 visits?

#### Solution:

- (a) Probability that the customer's first purchase of milk happens on 5<sup>th</sup> trip is given by

$$P(X = 5) = (1 - 0.2)^4 \times 0.2 = 0.08192$$

- (b) The average time between purchase of milk is

$$E(X) = \frac{1}{p} = \frac{1}{0.2} = 5$$

- (c) Given that a customer has not purchased milk during the past 3 shopping visits, the probability that the customer will not buy for another 2 visits is given by

$$P(X > 3+2 | X > 3) = P(X > 2) = (1-p)^2 = (1-0.2)^2 = 0.64$$

## 3.11 | PARAMETERS OF CONTINUOUS DISTRIBUTIONS

Continuous distributions are defined using the following three parameters:

1. **Scale parameter:** Scale parameter defines the range of the continuous distribution. The larger the scale parameter value, larger is the spread of the distribution.

2. **Shape parameter:** Shape parameter defines the shape of the probability distribution. The changes to the value of shape parameter will change the shape of the distribution.
3. **Location parameter:** Location parameter locates (or shifts) the distribution on the horizontal axis.

**Note:** Not all parametric distributions will have all three parameters.

### 3.12 | UNIFORM DISTRIBUTION

Uniform distribution is one of the simplest continuous distributions. Its probability density function and cumulative distribution functions are given by

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b] \\ 0, & \text{otherwise} \end{cases} \quad (3.34)$$

$$F(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \leq x < b \\ 1, & x \geq b \end{cases} \quad (3.35)$$

Mean and variance of uniform distribution are

$$E(X) = \frac{1}{2}(a+b) \text{ and } \text{Var}(X) = \frac{1}{12}(b-a)^2 \quad (3.36)$$

Uniform distribution is regularly used for gaining insights about a problem when other distributions are mathematically intractable for analysis in a given context.

### 3.13 | EXPONENTIAL DISTRIBUTION

Exponential distribution is a single parameter continuous distribution that is traditionally used for modelling time to failure of electronic components. The probability density function and cumulative distribution of exponential distribution are given by

$$f(x) = \lambda e^{-\lambda x}, \quad \lambda > 0 \quad (3.37)$$

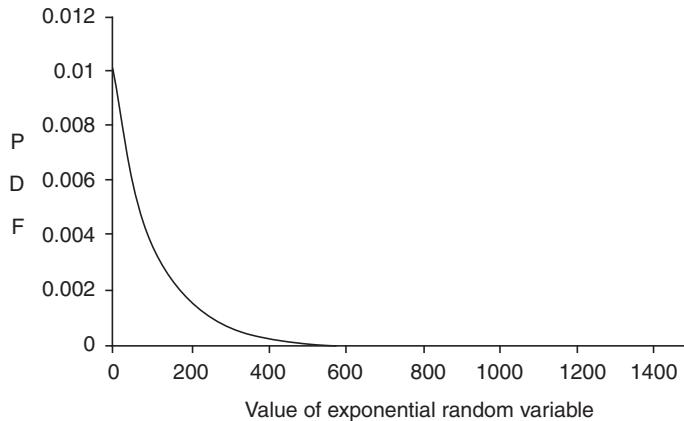
$$F(x) = 1 - e^{-\lambda x} \quad (3.38)$$

The parameter  $\lambda$  is the scale parameter and represents the rate of occurrence of the event,  $(1/\lambda)$  is the mean time between events. Probability density function and cumulative density function of an exponential distribution (with  $\lambda = 0.01$ ) are shown in Figures 3.13 and 3.14, respectively.

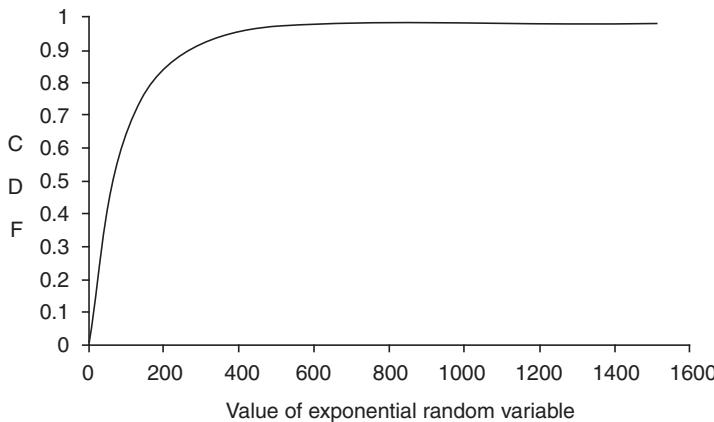
The mean and variance of an exponential distribution are given by

$$E(X) = \frac{1}{\lambda} \text{ and } \text{Var}(X) = \frac{1}{\lambda^2} \quad (3.39)$$

The expected value  $(1/\lambda)$  is the mean time between events.



**FIGURE 3.13** Probability density function of an exponential distribution.



**FIGURE 3.14** Cumulative distribution function of an exponential distribution.

### 3.13.1 | Memoryless Property of Exponential Distribution

Exponential distribution is the only continuous probability distribution that has the memoryless property. That is,

$$P(X > t + s \mid X > t) = P(X > s)$$

$$P(X > t + s \mid X > t) = \frac{P(X > t + s \cap X > t)}{P(X > t)} = \frac{P(X > t + s)}{P(X > t)} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} = e^{-\lambda s} \quad (3.40)$$

Memoryless property (also known as **Markov property**) of exponential distribution is exploited heavily in several analytics problems such as reliability analysis, warranty analytics, etc. Queueing models such as M/M/1 queues assume that the time between arrivals of customers follow an exponential distribution which has wide range of applications in capacity optimization under different business contexts.

**EXAMPLE 3.10**

The time to failure of an avionic system follows an exponential distribution with a mean time between failures (MTBF) of 1000 hours.

- Calculate the probability that the system will fail before 1000 hours.
- Calculate the probability that it will not fail up to 2000 hours.
- Calculate the time by which 10% of the systems will fail (that is, calculate  $P_{10}$  life).

**Solution:**

- The probability that the system will fail by 1000 hours is

$$F(1000) = 1 - e^{-\lambda t}$$

In this case,  $\lambda = 1/1000$ ,  $t = 1000$ . So

$$F(1000) = 1 - e^{-\frac{1}{1000} \times 1000} = 1 - e^{-1} = 0.6321$$

- The probability that the system will not fail up to 2000 hours is

$$P(X > 2000) = 1 - P(X \leq 2000) = 1 - F(t) = e^{-\lambda t} = e^{-\frac{1}{1000} \times 2000} = e^{-2} = 0.1353$$

- The time by which 10% of the systems will fail is

$$F(t) = 0.10 \Rightarrow 1 - e^{-\lambda t} = 0.1 \Rightarrow e^{-\lambda t} = 0.9$$

So

$$t = -\left(\frac{1}{\lambda}\right) \ln(0.9) = -1000 \times \ln(0.9) = 105.36 \text{ hours}$$

That is, by 105.36 hours, 10% of items will fail.

**EXAMPLE 3.11**

Waiting time at an airport checking counter follows an exponential distribution with mean time of 20 minutes.

- Calculate the probability that the waiting time is less than 5 minutes.
- If a customer has been waiting at the checking counter for 20 minutes, what is the probability that the customer will wait for 20 minutes more?

**Solution:**

- (a) The value of scale parameter  $\lambda = (1/20)$ . The probability the customer has to wait for less than 5 minutes is given by

$$F(5) = 1 - e^{-\frac{1}{20} \times 5} = 0.2211$$

- (b) If a customer has been waiting for 20 minutes, the probability that the customer will wait for additional 20 minutes is given by

$$P(X > 20 + 20 | X > 20) = P(X > 20) = e^{-\left(\frac{1}{20}\right) \times 20} = e^{-1} = 0.3678$$

In the above equation, we have used the memoryless property of exponential distribution.

### 3.14 | NORMAL DISTRIBUTION

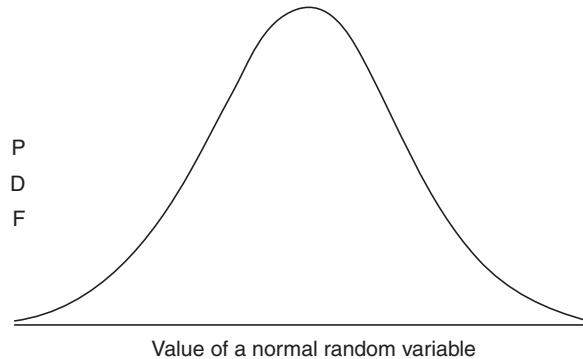
Normal distribution, also known as **Gaussian distribution**, is one of the most popular continuous distribution in the field of analytics especially due to its use in multiple contexts. Normal distribution is observed across many naturally occurring measures such as birth weight, height, intelligence, etc. The probability density function and the cumulative distribution function are given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < +\infty \quad (3.41)$$

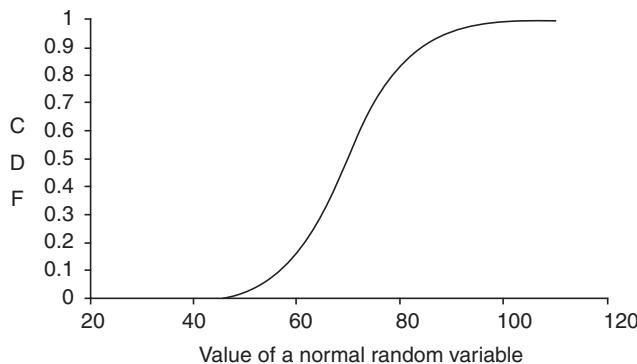
$$F(x) = \int_{-\infty}^x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt, \quad -\infty < x < +\infty \quad (3.42)$$

Here  $\mu$  and  $\sigma$  are the mean and standard deviation of the normal distribution. Normal distribution with mean  $\mu$  and standard deviation  $\sigma$  is denoted as  $N(\mu, \sigma^2)$ . Normal distribution is defined between  $-\infty$  and  $+\infty$ . For a normal distribution,  $\mu$  is the location parameter, which locates (center) the distribution on the horizontal axis and  $\sigma$  is the scale parameter, which defines the spread of the normal distribution. Normal distribution has no shape parameter since all normal distribution curves have bell shape and are symmetrical. Normal density curve and cumulative distribution curve are shown in Figures 3.15 and 3.16. In Microsoft Excel, the functions `NORM.DIST(x, μ, σ, false)` and `NORM.DIST(x, μ, σ, true)` can be used for calculating the probability density function and cumulative distribution function of a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ .

Historically, normal distribution was used in quantifying measurement errors associated with astronomical objects (Stahl, 2006). The error here is measured from the mean value and thus often called an *error function*. The curve can be also interpreted as ‘small errors are more frequent than large errors’. No closed form solution exists for the cumulative distribution function of a normal distribution. Many functions that are numerical approximations are used for finding the value of cumulative distribution function of normal distribution.



**FIGURE 3.15** Probability density function of a normal distribution.



**FIGURE 3.16** Cumulative distribution function of a normal distribution.

### 3.14.1 | Properties of Normal Distribution

1. Theoretical normal density functions are defined between  $-\infty$  and  $+\infty$ .
2. It is a two parameter distribution, where the parameter  $\mu$  is the mean (location parameter) and the parameter  $\sigma$  is the standard deviation (scale parameter).
3. All normal distributions have symmetrical bell shape around mean  $\mu$  (thus it is also median).  $\mu$  is also the mode of the normal distribution, that is,  $\mu$  is the mean, median as well as the mode.
4. For any normal distribution, the areas between specific values measured in terms of  $\mu$  and  $\sigma$  are given by:

Value of Random Variable	Area under the Normal Distribution (CDF)
$\mu - \sigma \leq X \leq \mu + \sigma$ (area between one sigma from the mean)	0.6828
$\mu - 2\sigma \leq X \leq \mu + 2\sigma$ (area between two sigma from the mean)	0.9545
$\mu - 3\sigma \leq X \leq \mu + 3\sigma$ (area between three sigma from the mean)	0.9973

5. Any linear transformation of a normal random variable is also normal random variable. That is, if  $X$  is a normal random variable, then the linear transformation  $AX + B$  (where  $A$  and  $B$  are two constants) is also a normal random variable.
6. If  $X_1$  and  $X_2$  are two independent normal random variables with mean  $\mu_1$  and  $\mu_2$  and variance  $\sigma_1^2$  and  $\sigma_2^2$ , respectively, then  $X_1 + X_2$  is also a normal distribution with mean  $\mu_1 + \mu_2$  and variance  $\sigma_1^2 + \sigma_2^2$ .
7. Sampling distribution of mean values of a large sample drawn from a population of any distribution is likely to follow a normal distribution. This result is known as the *central limit theorem* and will be discussed in detail in Chapter 4.

### 3.14.2 | Standard Normal Variable

A normal random variable with mean  $\mu = 0$  and  $\sigma = 1$  is called the standard normal variable and usually represented by  $Z$ . The probability density function and cumulative distribution function of a standard normal variable are given by

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \quad (3.43)$$

$$F(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \quad (3.44)$$

By using the following transformation, any normal random variable  $X$  can be converted into a standard normal variable:

$$Z = \frac{X - \mu}{\sigma} \quad (3.45)$$

The random variable  $X$  can be written in the form of a standard normal random variable using the relationship

$$X = \mu + \sigma Z \quad (3.46)$$

Thus, any normal random variable  $X$  can be expressed using the standard normal random variable  $Z$ . No closed form solution exists for the cumulative standard normal distribution; however, there are several approximate formulas available for calculating CDF of a standard normal distribution (Yerukala and Boiroju, 2015). A simple approximation of standard normal CDF is given by Tocher (1963) as follows:

$$P(Z \leq z) = F(z) \approx \frac{e^{2kz}}{1 + e^{2kz}} \quad (3.47)$$

where  $k = \sqrt{2/\pi}$ .

Another more accurate approximation is provided by Byrc (2002):

$$P(Z \leq z) = F(z) = 1 - \left( \frac{z^2 + A_1 z + A_2}{\sqrt{2\pi} \times z^3 + B_1 z^2 + B_2 z + 2A_2} \right) \times e^{-z^2/2} \quad (3.48)$$

where  $A_1$ ,  $A_2$ ,  $B_1$ , and  $B_2$  are constants given by

$$A_1 = 5.575192695; A_2 = 12.77436324; B_1 = 14.38718147; B_2 = 31.53531977$$

Note that, any normal random variable can be converted into standard normal random variable and the above approximations can be used for finding the CDF value.

In Microsoft Excel®, the normal PDF value is given by the function Normdist( $x, \mu, \sigma, \text{false}$ ) and CDF value is given by Normdist( $x, \mu, \sigma, \text{true}$ ), where  $x$  is the value of the normal random variable,  $\mu$  is the mean of the distribution, and  $\sigma$  is the corresponding standard deviation. The CDF value of a standard normal distribution can be obtained using the function NORMSDIST( $Z$ ) in Microsoft Excel.

### EXAMPLE 3.12

According to a survey on use of smart phones in India, the smart phone users spend 68 minutes in a day on average in sending messages and the corresponding standard deviation is 12 minutes. Assume that the time spent in sending messages follows a normal distribution.

- (a) What proportion of the smart phone users are spending more than 90 minutes in sending messages daily?
- (b) What proportion of customers are spending less than 20 minutes?
- (c) What proportion of customers are spending between 50 minutes and 100 minutes?

**Solution:**

It is given that  $\mu = 68$  minutes and  $\sigma = 12$  minutes.

- (a) Proportion of customers spending more than 90 minutes is given by

$$P(X \geq 90) = 1 - P(X \leq 90) = 1 - F(90)$$

The standard normal random variable value for  $X = 90$  is given by

$$Z = \frac{x - \mu}{\sigma} = \frac{90 - 68}{12} = 1.8333$$

That is,  $F(X = 90) = F(Z = 1.8333)$ . From standard normal distribution table, we can get the value of  $F(Z)$  for  $Z = 1.8333$ . The area under the standard normal distribution curve for  $Z = 1.8333$  is 0.9666. Thus,

$$P(X \geq 90) = 1 - P(X \leq 90) = 1 - F(90) = 1 - 0.9666 = 0.0334$$

Alternatively, using Excel, we get

$$P(X \geq 90) = 1 - P(X \leq 90) = 1 - \text{Normdist}(90, 68, 12, \text{true}) = 0.0334$$

- (b) Proportion of customers spending less than 20 minutes is

$$P(X \leq 20) = F(20)$$

Using Excel function, we have  $\text{Normdist}(20, 68, 12, \text{true}) = 3.1671 \times 10^{-5}$

- (c) Proportion of customers spending between 50 and 100 minutes is given by

$$\begin{aligned} P(50 \leq X \leq 100) &= F(100) - F(50) \\ &= \text{Normdist}(100, 68, 12, \text{true}) - \text{Normdist}(50, 68, 12, \text{true}) \\ &= 0.9293 \end{aligned}$$

### EXAMPLE 3.13

At Die Another Day (DAD) hospital, nurses are given an additional bonus of INR 1,00,000 if they stay for more than 36 months with DAD hospital. The average stay of nurses follows a normal distribution with an average of 28 months and the corresponding standard deviation is 4.8 months. Calculate

- (a) The expected number of nurses who will be given bonus and the value of bonus that will be given if 50 new nurses join DAD hospital in the current month,
- (b) What will be the additional amount paid if DAD hospital changes the policy that they will give bonus if the stay exceeds 24 months? What assumptions are made in this case?

**Solution:**

- a) Expected number of nurses and the value of bonus:

$$\text{Expected number of nurses who will be getting bonus} = 50 \times P(X \geq 36)$$

$$P(X \geq 36) = 1 - \text{Normdist}(36, 28, 4.8, \text{true}) = 0.04779$$

$$\text{Expected number of nurses who will be getting bonus} = 50 \times 0.04779 = 2.389518$$

$$\text{Expected value of bonus given} = 50 \times P(X \geq 36) \times 100,000 = \text{INR } 238951.76$$

- (b) The additional bonus given is

$$50 \times 1,00,000 \times [\text{Normdist}(36, 28, 4.8, \text{true}) - \text{Normdist}(24, 28, 4.8, \text{true})] = 3749406$$

The major assumption here is that the policy change is unlikely to change the attrition behaviour of the nurses, which may not be true. Since the nurses now know that if they stay for 24 months, they will get the bonus, the distribution parameter values are likely to change.

### 3.15 | CHI-SQUARE DISTRIBUTION

Chi-square distribution with  $k$  degrees of freedom [denoted as  $\chi^2(k)$  distribution] is a non-parametric distribution which is obtained by adding square of  $k$  independent standard normal random variables. Chi-square distribution plays a pivotal role in analytics since it is used in many hypothesis tests and forms the basis for the classification tree technique chi-square automatic interaction detection (CHAID). In sentiment analysis, chi-square test of independence is used to identify the features (words) that can be used for classifying documents into various classes such as positive, negative, and neutral. In logistic regression model, chi-square test is used to find the statistical significance of independent variables (predictor variables).

Consider a normal random variable  $X_1$  with mean  $\mu_1$  and standard deviation  $\sigma_1$ . Then we can define  $Z_1$  (the standard normal random variable) as

$$Z_1 = \frac{X_1 - \mu_1}{\sigma_1}$$

Then,

$$Z_1^2 = \left( \frac{X_1 - \mu_1}{\sigma_1} \right)^2 \quad (3.49)$$

is a chi-square distribution with one degree of freedom [ $\chi^2(1)$ ]. Let  $X_2$  be a normal random variable with mean  $\mu_2$  and standard deviation  $\sigma_2$  and  $Z_2$  be the corresponding standard normal variable. Then the random variable  $Z_1^2 + Z_2^2$  given by

$$Z_1^2 + Z_2^2 = \left( \frac{X_1 - \mu_1}{\sigma_1} \right)^2 + \left( \frac{X_2 - \mu_2}{\sigma_2} \right)^2 \quad (3.50)$$

is a chi-square distribution with 2 degrees of freedom. A chi-square distribution with  $k$  degrees of freedom is given by sum of squares of standard normal random variables  $Z_1, Z_2, \dots, Z_k$  obtained by transforming normal random variables  $X_1, X_2, \dots, X_k$  with mean values  $\mu_1, \mu_2, \dots, \mu_k$  and corresponding standard deviations  $\sigma_1, \sigma_2, \dots, \sigma_k$ . That is

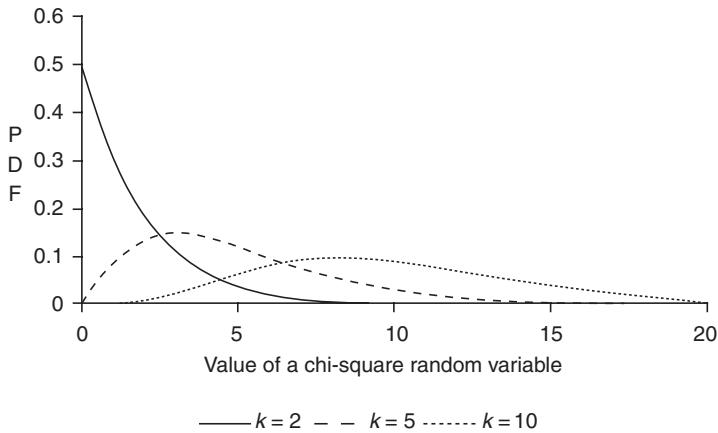
$$\chi^2(k) = Z_1^2 + Z_2^2 + \dots + Z_k^2 = \left( \frac{X_1 - \mu_1}{\sigma_1} \right)^2 + \left( \frac{X_2 - \mu_2}{\sigma_2} \right)^2 + \dots + \left( \frac{X_k - \mu_k}{\sigma_k} \right)^2 \quad (3.51)$$

The probability density function of  $\chi^2(k)$  is given by

$$f(x) = \frac{1}{2^{k/2} \Gamma(k/2)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}} \quad (3.52)$$

where  $\Gamma(k/2)$  is a Gamma function given by

$$\Gamma(k) = \int_0^\infty x^{k-1} e^{-x} dx \quad (3.53)$$



**FIGURE 3.17** Probability density function of chi-square distribution for different values of  $k$ .

The probability density function of  $\chi^2(k)$  for different values of  $k$  is shown in Figure 3.17.

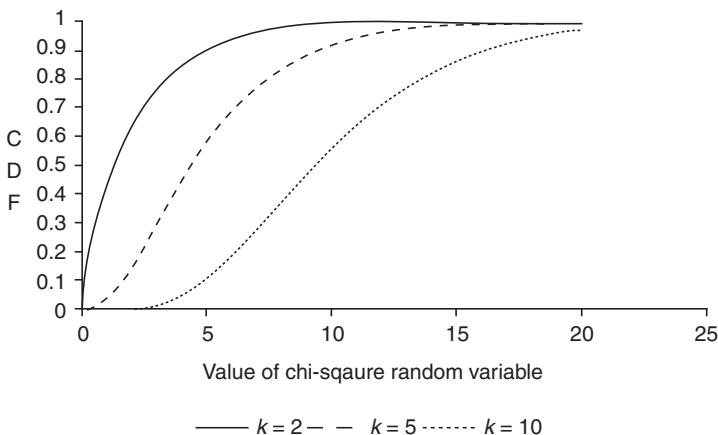
The cumulative distribution function of a chi-square distribution with  $k$  degrees of freedom is given by

$$F(x) = \frac{\gamma\left(\frac{k}{2}, \frac{x}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \quad (3.54)$$

where  $\gamma\left(\frac{k}{2}, \frac{x}{2}\right)$  is the lower incomplete Gamma function. It is given by

$$\gamma(k, x) = \int_0^x t^{k-1} e^{-t} dt \quad (3.55)$$

Cumulative chi-square distribution for different value of  $k$  is shown in Figure 3.18.



**FIGURE 3.18** Cumulative distribution of chi-square distribution with  $k$  degrees of freedom.

In Microsoft Excel, the function ‘CHISQ.DIST( $x, k, \text{true}$ )’ can be used for calculating the cumulative distribution function value of chi-square distribution with  $k$  degrees of freedom.

### 3.15.1 | Properties of Chi-Square Distribution

1. The mean and standard deviation of a chi-square distribution are  $k$  and  $\sqrt{2k}$  respectively, where  $k$  is the degrees of freedom.
2. As the degrees of freedom  $k$  increases, the probability density function of a chi-square distribution approaches normal distribution.
3. Chi-square goodness of fit test is one of the popular tests for checking whether a data follows a specific probability distribution.

## 3.16 | STUDENT'S $t$ -DISTRIBUTION

Student's  $t$ -distribution (or simply  $t$ -distribution) arises while estimating the population mean of a normal distribution using sample which is either small and/or the population standard deviation is unknown. The distribution was developed by William Gosset under the pseudo name 'student' while working for *Guinness Brewery* in Dublin, Ireland (Student, 1908) and thus is called student's distribution (since the letter  $t$  was used for representing the distribution it is called student's  $t$ -distribution). In analytics,  $t$ -distribution plays an important role in hypothesis testing and also in diagnostics of linear regression models.

Assume that  $X_1, X_2, \dots, X_n$  are  $n$  observations (that is, sample of size  $n$ ) from a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ . Let

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

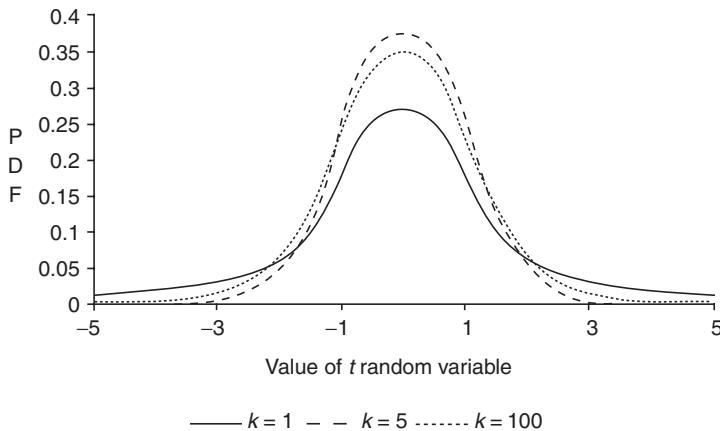
$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$

where  $\bar{X}$  and  $S$  are mean and standard deviation estimated from the sample  $X_1, X_2, \dots, X_n$ . Then the random variable  $t$  defined by

$$t = \frac{\bar{X} - \mu}{S / \sqrt{n}} \quad (3.56)$$

follows a  $t$ -distribution with  $(n - 1)$  degrees of freedom. Here one degree of freedom is lost since the standard deviation is estimated from the sample (degrees of freedom is the number of observations in the sample minus number of restrictions or estimates made using the sample). The probability density function of  $t$ -distribution with  $n$  degrees of freedom is given by

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right) \sqrt{n\pi}} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} \quad (3.57)$$



**FIGURE 3.19** Probability density function of student's  $t$ -distribution.

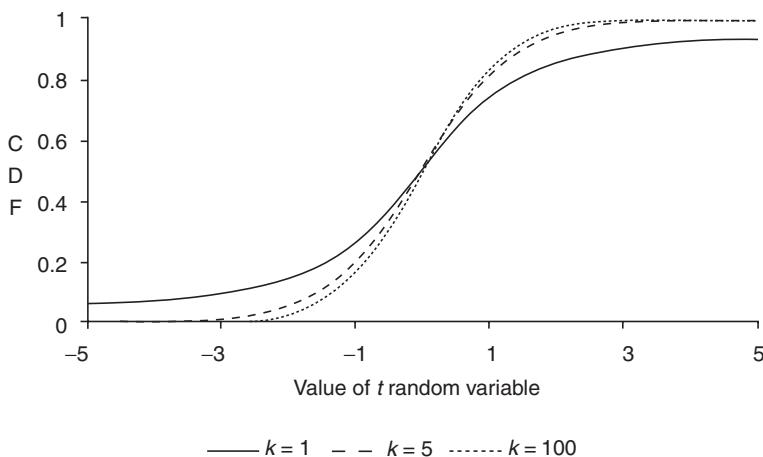
The probability density function of  $t$ -distribution for different values of  $n$  (degrees of freedom) is shown in Figure 3.19.

Cumulative distribution of student's  $t$ -distribution for different value of degrees of freedom is shown in Figure 3.20. In Microsoft Excel, the function 'T.DIST( $x$ ,  $n$ , true)' gives the cumulative distribution function value of  $t$ -distribution with  $n$  degrees of freedom.

An alternative definition of  $t$ -distribution is that, if  $Z$  is a standard normal distribution and  $\chi^2(k)$  is a chi-square distribution with  $k$  degrees of freedom, then

$$t = \frac{Z}{\sqrt{\frac{\chi^2(k)}{k}}} \quad (3.58)$$

is a  $t$ -distribution.



**FIGURE 3.20** Cumulative distribution function of student's  $t$ -distribution.

### 3.16.1 | Properties of t-Distribution

1. The mean of a  $t$ -distribution with 2 or more degrees of freedom is 0.
2. The variance of  $t$ -distribution is  $n/(n-2)$  for  $n > 2$ , where  $n$  is the number of degrees of freedom.
3. As the degrees of freedom  $n$  increases, the probability density function of a  $t$ -distribution approaches the density function of standard normal distribution. For  $n > 120$ , the difference between the area under probability density function of a  $t$ -distribution is very close to the area under a standard normal distribution.
4.  $t$ -distribution is an important distribution for hypothesis testing of means of a population and for comparing means of two populations.

### 3.17 | F-DISTRIBUTION

$F$ -distribution (short form of Fisher's distribution named after statistician Ronald Fisher) is a ratio of two chi-square distributions. Let  $Y_1$  and  $Y_2$  be two independent chi-square distributions with  $k_1$  and  $k_2$  degrees of freedom, respectively. Then the random variable  $X$  defined as (Abramowitz and Stegun, 1972)

$$X = \frac{Y_1 / k_1}{Y_2 / k_2} \quad (3.59)$$

is an  $F$ -distribution. The probability density function of an  $F$ -distribution is given by (source: Engineering Statistics Handbook)<sup>1</sup>

$$f(x) = \frac{\Gamma\left(\frac{k_1 + k_2}{2}\right)\left(\frac{k_1}{k_2}\right)^{k_1/2}}{\Gamma\left(\frac{k_1}{2}\right)\Gamma\left(\frac{k_2}{2}\right)} \times \frac{x^{\frac{k_1}{2}-1}}{\left(1 + \frac{k_1 x}{k_2}\right)^{\frac{k_1+k_2}{2}}} \quad (3.60)$$

The probability density function of an  $F$ -distribution for different values of  $k_1$  and  $k_2$  are shown in Figure 3.21. The cumulative distribution function is shown in Figure 3.22.

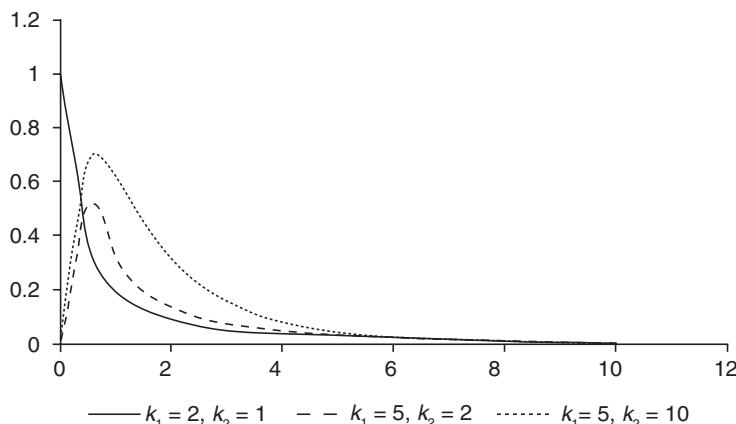


FIGURE 3.21 Probability density function of  $F$ -distribution.

<sup>1</sup> <http://www.itl.nist.gov/div898/handbook/eda/section3/eda3665.htm>

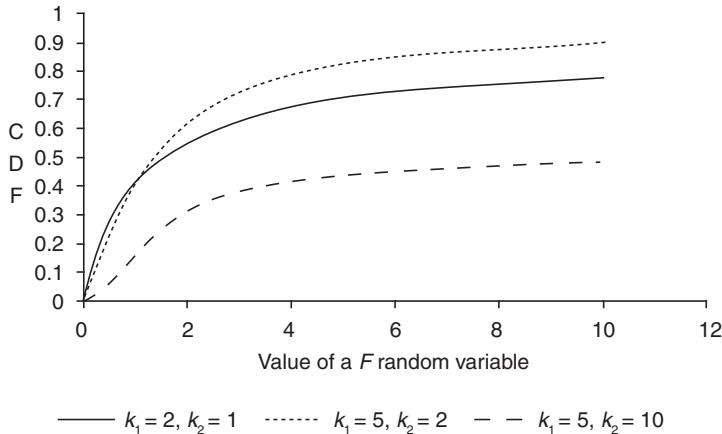


FIGURE 3.22 Cumulative distribution function of  $F$ -distribution.

$F$ -distribution is used in Analysis of Variance (ANOVA) which also plays an important role in diagnostics of overall fitness of a multiple linear regression model.

### 3.17.1 | Properties of $F$ -Distribution

- Mean of  $F$ -distribution is  $k_2 / (k_2 - 2)$ , for  $k_2 > 2$ .
- Standard deviation of  $F$ -distribution is  $\sqrt{\frac{2k_2^2(k_1 + k_2 - 2)}{k_1(k_2 - 2)^2(k_2 - 4)}}$  for  $k_2 > 4$ .
- $F$ -distribution is non-symmetrical and the shape of the distribution depends on the values of  $k_1$  and  $k_2$ .
- $F$ -distribution is used in Analysis of Variance to test the mean values of multiple groups (discussed in Chapter 7).

#### SUMMARY

- The concept of probability, random variables and probability distributions are foundations of data science. Knowledge of these concepts is important for framing and solving analytics problems.
- Random variable is a function that maps an outcome of a random experiment to a real number and plays an important role in analytics since many key performance indicators used across industries are random variables.
- Basic probability concepts such as joint events, independent events, conditional probability and Bayes' theorem are useful for predicting probability of an event of importance. These concepts are used in algorithms such as association rule learning which is used in solving analytics problems such as market basket analysis and recommender systems.
- Discrete probability distributions such as binomial distribution, Poisson distribution and geometric distribution are used for modelling discrete random variables.
- Continuous distributions such as normal distribution, chi-square distribution,  $t$ -distribution and  $F$ -distribution play an important role in hypothesis testing.

### MULTIPLE CHOICE QUESTIONS

1. If two events  $A$  and  $B$  are dependent, then
 

(a) $P(A \cap B) = P(A) + P(B) - P(A \cup B)$	(b) $P(A \cap B) = P(A) + P(B)$
(c) $P(A \cap B) = P(A) \times P(B A)$	(d) $P(A \cap B) = P(B) \times P(A B)$
2. Which of the following statements are correct?
  - (a) If the events are independent then they are also mutually exclusive.
  - (b) If the events are mutually exclusive then they are dependent events.
  - (c) If the events are independent events, then they cannot be mutually exclusive.
  - (d) If the events are mutually exclusive then they are independent.
3. Which of the following random variables are discrete random variables?
 

(a) Waiting time at an ATM	(b) Number of customers waiting at an ATM
(c) Time between failures of ATM	(d) Number of cash withdrawals at an ATM
4. An e-tailer (e-commerce retailer) is interested in predicting the number of returns that the e-tailer is likely to receive on any given day. Which of the following distribution is more appropriate for modelling this problem?
 

(a) Binomial distribution	(b) Normal distribution
(c) Poisson distribution	(d) Geometric distribution
5. Which of the following distributions do not have memoryless property?
 

(a) Geometric distribution	(b) Exponential distribution
(c) $t$ -distribution	(d) Normal distribution
6. If  $X$  is a random variable that follows normal distribution, then  $5X + 10$  is a
 

(a) Chi-square distribution	(b) Exponential distribution
(c) Normal distribution	(d) $t$ -distribution
7. Which of the following distributions are not related to normal distribution?
 

(a) $F$ -distribution	(b) Exponential distribution
(c) Chi-square distribution	(d) $t$ -distribution
8. If  $Z_1$  and  $Z_2$  are two standard normal random variables, then  $Z_1^2 + Z_2^2$  will be
 

(a) Exponential distribution	(b) $t$ -distribution
(c) $F$ -distribution	(d) Chi-square distribution
9. Which of the following distributions do not have shape parameter?
 

(a) Normal distribution	(b) Exponential distribution
(c) $F$ -distribution	(d) $t$ -distribution
10.  $X_1$  is a normal distribution with mean 500 and standard deviation 50 and  $X_2$  is a normal distribution with mean 800 and standard deviation 100. Then
 

(a) $P(X_1 \leq 450) = P(X_2 \leq 700)$	(b) $P(X_1 \leq 450) > P(X_2 \leq 700)$
(c) $P(X_1 \leq 450) < P(X_2 \leq 700)$	(d) $P(X_1 \leq 450) \neq P(X_2 \leq 700)$

### EXERCISES

1. Bank of Palakkad (BoP) provides two types of loans: (a) gold loans and (b) personal loans. 90% of the all loans at BoP are gold loans and the rest are personal loans. The non-performing asset (NPA) at BoP is 6%. The joint probability of gold loan and NPA is 4%. If a customer is given a gold loan, calculate the probability that this loan will become an NPA.
2. The Indian market is classified into 4 regions: (a) North, (b) South, (c) East, and (d) West by an e-tailer. The market share of 4 regions are 40%, 30%, 10%, and 20%, respectively. The percentages of returns from

customers for north, south, east, and west are 12%, 4%, 7%, and 10%, respectively. If the e-tailer receives a return from its customer, calculate the probability that the return is from the region south.

3. Local Dhania (LD) an online grocery store segments customers into three categories (say A, B, and C). The percentage of customers in segments A, B, and C are 70%, 20%, and 10%, respectively. LD regularly collects feedback from their customers using an 11 point scale (0 to 10). Customers providing a rating between 0 and 6 are labelled as detractors, customers giving a feedback score of 7 and 8 are classified as passives, and those giving feedback score of 9 and 10 are classified as promoters. The percentage of detractors from segments A, B, and C are 4%, 6%, and 9%, respectively. Based on feedback, a customer is classified as detractor. Calculate the probability that this customer belongs to segment C.
4. An airline operates several domestic flights from 4 major airports of India: (a) New Delhi, (b) Mumbai, (c) Bangalore, and (d) Hyderabad. The percentage of flights operated from New Delhi, Mumbai, Bangalore, and Hyderabad are 40%, 25%, 25%, and 10%, respectively. The percentage of delayed flights at these four airports are 10%, 8%, 7%, and 6%. If a flight is delayed, what is the probability that the flight originated from Bangalore airport?
5. Software India Inc (SII) is an IT company based out of Bangalore. A frequent problem they encounter while recruiting new staff is that few candidates accept the job offer but do not join (renege). The percentage of candidates who renege is 20%. If SII makes 8 offers:
  - (a) What is the probability that all 8 candidates will join after accepting the offer?
  - (b) What is the probability exactly 2 candidates will not join?
6. At Die Another Day (DAD) hospital, the probability of a patient complaining about the service is 0.05.
  - (a) If 10 patients are admitted, what is the probability that none of them will complain about the service?
  - (b) What is the probability at least one of them will complain?
7. The proportion of customers who abandon the cart at an e-commerce site is 15%. Every day 25000 customers visit the e-commerce site and add items to their cart. What is the probability that at least 1000 customers will abandon the cart? (Hint: Use normal approximation for binomial distribution.)
8. The arrival of customers at an automatic teller machine (ATM) follows a Poisson distribution and the average number of arrivals is 20 per hour.
  - (a) What is the probability that exactly 10 customers will arrive in an hour?
  - (b) What is the probability that more than 15 customers will arrive in one hour?
9. The number of warranty claims received at an auto manufacturer follows a Poisson distribution with a rate of 12 per day. Calculate
  - (a) The probability that the warranty claims will exceed 100 over 10 days.
  - (b) The average number of claim over 10 days.
10. A machine learning algorithm is used to classify images; the probability of error in classification is 0.01.
  - (a) What is the average number of correct predictions before the algorithm makes an incorrect prediction?
  - (b) What is the probability that the first error occurs after exactly 10 correct predictions?
  - (c) What is the probability that the algorithm makes at least 10 correct predictions before an incorrect prediction?
11. The probability of bowling a wide ball by a bowler in a cricket match is 0.10. Calculate the following:
  - (a) Probability that a bowler's 5<sup>th</sup> ball in a match will be wide ball.
  - (b) In a T20 match, each bowler can bowl a maximum of 4 overs (24 legal deliveries). What is the probability that a bowler will not bowl a wide ball during his entire spell of 4 overs?
  - (c) If the bowler bowls 60 balls (legal deliveries), calculate the expected number of wide balls.
12. The waiting time at the airport security check follows an exponential distribution with a mean of 20 minutes. Ms Thennal K Warrier arrives at the security check 40 minutes prior to the departure of her flight. It

takes 5 minutes to reach the boarding gate after clearing the security check and the airline will close the gate 10 minutes prior to the departure of the flight. Passengers not reaching the gate 10 minutes prior to departure are not allowed to board the flight.

- (a) What is the probability that Ms Thennal K Warrier will miss the flight?
  - (b) What is the probability that Ms Thennal K Warrier will wait more than 40 minutes at the security check?
  - (c) Ms Thennal K Warrier has been waiting for 20 minute at the security gate. What is the probability that she will wait for another 20 minutes?
  - (d) If Ms Thennal K Warrier would like to ensure that she would not like to miss the flight in 99% cases, how many hours before the flight departure should she reach the airport security?
13. Time to failure distribution of an auxiliary power unit follows an exponential distribution with mean time between failures of 2000 hours.
- (a) What is the probability that the auxiliary power unit will survive for at least 1000 hours?
  - (b) The manufacturer of the auxiliary power unit would like to decide on the warranty such that not more than 5% of the auxiliary power unit can fail during the warranty period. What should be the duration of warranty?
14. Local Dhania, the online grocery store, makes a promise that it will deliver the order within 90 minutes. Based on the past data, it was found that the average time taken to deliver is 68 minutes and the corresponding standard deviation is 14 minutes, and follows a normal distribution.
- (a) What proportion of orders is delivered after 90 minutes?
  - (b) If LD would like to ensure that at least 99% of the orders are delivered before 90 minutes, what should be the target mean delivery time (assume that there is no change in the standard deviation)?
  - (c) If it is not possible to reduce the mean time to deliver and standard deviation, what should be the promised time to deliver (instead of current 90 minutes) for which the probability of delivery before that time is 99%?
15. The patient registration time at a hospital follows a normal distribution with a mean of 12 minute and a standard deviation of 3 minutes. Every week, approximately 800 new patients are admitted who are required to register first. How many of these patients will be waiting for more than 10 minute at the registration desk? To improve customer satisfaction, the hospital introduces a policy that every patient who is made to wait for more than 5 minutes at the registration desk is given a free meal coupon. How many free coupons are given by the hospital every week?
16. A random sample of size 100 is taken from a normal population and sample mean and standard deviation are estimated as 250 and 40, respectively. Calculate the probability that the population mean is more than 260.

## REFERENCES

1. Abramowitz M and Stegun I A (1972), *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables* (9<sup>th</sup> Edition), Dover, New York.
2. Bryc W (2002), “A Uniform Approximation to the Right Normal Tail Integral”, *Applied Mathematics and Computation*, **127**, 365–374.
3. Kolmogorov A (1956), *Foundation of the Theory of Probability (Translation)*, Chelsea Publishing Company, New York.
4. Rao S S (1998), “Diaper-beer Syndrome”, Forbes, 4 June 1998. Available at <https://www.forbes.com/forbes/1998/0406/6107128a.html>, accessed on 20 March 2017.
5. Stahl S (2006), “The Evolution of Normal Distribution”, *Mathematics Magazine*, **79**(2), 96–113.
6. Student (1908), “The Probable Error of Means”, *Biometrika*, **6**, 1–25.
7. Toucher K D (1963), *The Art of Simulation*, English University Press, London.
8. Yerukala R and Boiroju N K (2015), “Approximations to Standard Normal Distribution Function”, *International Journal of Scientific and Engineering Research*, **6**(4), 515–518.

# 4

# Sampling and Estimation

“To Clarify \*add\* data.”

— Edward R Tufte

## LEARNING OBJECTIVES

- LO 4-1** Understand the need for sampling and the importance of appropriate sampling.
- LO 4-2** Understand the difference between population parameters and sample statistic.
- LO 4-3** Learn different types of sampling techniques and limitations of each sampling approach.
- LO 4-4** Learn about estimation of parameters and sampling distribution.
- LO 4-5** Understand Central Limit Theorem (CLT) and its importance in hypothesis testing.
- LO 4-6** Learn method of moments and maximum likelihood estimator (MLE) and its applications in estimation of parameters of probability distributions.

## ESSENCE OF SAMPLING

Sampling is a process of selecting subset of observations from a population to make inference about various population parameters such as mean, proportion, standard deviation, etc. It is an important step in inferential statistics since an incorrect sample may lead to wrong inference about the population. Sampling process itself has several steps and each step is important to ensure that the ideal sample is used for estimation of population parameters and for making inferences about the population. Under Big Data context, we may use almost the entire population; however, in most cases we will still be dependent on samples to make inference.

IMPORTANT

*Sampling is necessary when it is difficult or expensive to collect data on the entire population. The inference about the population is made based on the sample that was collected; incorrect sample may lead to incorrect inference about the population.*

## 4.1 | INTRODUCTION TO SAMPLING

The population of India was close to 1.32 billion in July 2016 according to United Nations (Source: Wikipedia<sup>1</sup>). Census India collects information such as demography, literacy, housing, economic activity of the individuals, etc. Thousands of people are used for collecting data on such huge population,

<sup>1</sup> Source: [https://en.wikipedia.org/wiki/Demographics\\_of\\_India](https://en.wikipedia.org/wiki/Demographics_of_India)

which is very expensive and thus carried out only once in 10 years. Many organizations cannot afford to collect data on the entire population and in many cases it is expensive and time-consuming. In many cases, all members of the population are not known for sampling purpose. For example, assume that an organization is interested to conduct a study on diabetic patients. According to Lancet, the estimated diabetic patients in India in 2014 were 64.5 million (Mascarenhas, 2016). It is impossible to collect data from all diabetic patients, since many diabetic patients themselves may not be aware that they are diabetic, especially during early stages. Even when all the members of the population are known, collecting data can be expensive and thus in many cases we have to settle for samples. Sampling is necessary because even when the entire population is available, using the entire population for estimation of a population parameter may not be feasible. Consider reliability estimation of engineering systems, the original equipment manufacturer (OEM) may carry out some destructive testing (in which the item is destroyed in the test, for example crash testing of cars) to understand the causes of failure and estimate the population parameter such as mean time between failure (MTBF); for obvious reasons OEM cannot use the entire population of systems for destructive testing.

Sampling is an important activity in analytics, especially when inferences are made about the population based on the sample. Incorrect sampling can result in incorrect estimation of population parameters and wrong inference about the population. One of the frequently used examples to demonstrate the importance of sampling is the opinion poll conducted by the Literary Digest in 1936 USA presidential election (Squire, 1988). Alfred Landon, the Republican Governor of Kansas, was contesting against the incumbent Franklin D Roosevelt. Literary Digest predicted that Landon would get 55%, Roosevelt 41%, and Lemke 4% votes. However, the actual votes polled were 61% for Roosevelt and 37% for Landon, resulting in a prediction error of 20% for Roosevelt (Squire, 1988). The main reason for such a huge error was attributed to the sampling method employed by Literary Digest, in spite of the fact that their sample size was about 2.4 million. There were two major issues with the sampling processes used by Literary Digest. The first one was the selection of voters for the poll (sampling frame); the names were taken from telephone directories, subscribers of the magazines, club members, automobile registry, etc. (Squire, 1988). In 1936, this meant that they were selecting middle and upper middle class voters since telephone ownership was a rarity in 1936. Thus, the sampling framework used had an in-built selection bias of individuals for the purpose of study. The second was that they contacted close to 10 million voters; however, only 2.4 million responded (which in itself is a very large sample). But such low response can result in non-responsive bias. Cahalan (1989) reported that based on his survey 67% Roosevelt supporters and 52% Landon supporters claimed that they had not received Literary Digest ballot. Literary Digest went bankrupt after this prediction. The election result was correctly predicted by George Gallup, a pioneer of survey sampling with just 50,000 samples (Squire, 1988). This example clearly demonstrates the importance of selection of items in the sample and a large sample does not improve prediction if the sampling process is incorrect. Even in the age of analytics and big data, the winner of the 2016 US presidential election was incorrectly predicted by media. Most media outfits predicted comfortable win for Ms Hillary Clinton. Most of these incorrect predictions can be attributed to incorrect sampling procedures (mostly biased data collection and even biased reporting).

#### 4.1.1 | When Jesus Christ Became 2<sup>nd</sup> Best in the World

Respondent bias is another source problem in survey sampling. There was an internet poll conducted in 1998 to find the ‘most influential figure of the last 2000 years’. Jamie Pollock who played as defensive midfielder for the football team Manchester City won the title leaving Jesus Christ and Carl Marx to second and third place, respectively (Szczepanik, 2016). Apparently the poll was rigged by the supporters of the club Queens Park Rangers (QPR) who voted multiple times for Jamie Pollock. The story goes like this: On 25<sup>th</sup> April 1998, Manchester City was playing against QPR in Division 1 of English football and both teams were in danger of being relegated to division 2 (third tier of English football). When the goal was 1–1, Jamie Pollock scored an own goal giving upper hand to QPR. Although the match ended in a 2–2 draw, Manchester City was relegated to Division 2 for the first time and QPR managed stay in Division 1 (Moxley 2012, Szczepanik 2016). Collecting survey and ensuring that the sample is unbiased is one of the major challenges in analytics. Summers (1969) grouped bias in survey research in to several categories such as (a) sampling bias, (b) non-responsive bias, (c) respondent bias, (d) instrumentation bias, etc.

## 4.2 | POPULATION PARAMETERS AND SAMPLE STATISTIC

In many real-life problems, the population can be very large making it impossible to collect every feature of each case in the population. Measures such as mean and standard deviation calculated using the entire population are called *population parameters*. The population parameters mean and standard deviation are usually denoted using symbols  $\mu$  and  $\sigma$ , respectively. Since calculating population parameters in most practical situations is almost impossible, we depend on samples to estimate the population parameters. Population parameters estimated from sample are called *sample statistic* or *statistic*. The sample statistic is denoted using symbols  $\bar{X}$  (for mean) and  $S$  (or  $s$  for standard deviation). Statisticians also use hat symbol ( $\hat{X}$  for mean and  $\hat{\sigma}$  for standard deviation) for statistic. Since inferences about population are made using a sample, statistic plays an important role in hypothesis testing.

In addition to sample mean and standard deviation, we frequently estimate population proportion ( $p$ ), which is proportion of cases in the data belonging to a specific category. Assume that a bank classifies its customers based on the risk categories: (1) Low, (2) Medium, and (3) High. We would like to know what proportions of the population belong to categories 1, 2, and 3 which are denoted by  $p_1$ ,  $p_2$ , and  $p_3$ , respectively. The corresponding estimates will be denoted by  $\hat{p}_1$ ,  $\hat{p}_2$  and  $\hat{p}_3$ . Assume that  $n_1$ ,  $n_2$ , and  $n_3$  are the number of cases under categories 1 (low risk), 2 (medium risk), and 3 (high risk), respectively. The estimates of proportions are given by

$$\hat{p}_1 = \frac{n_1}{n_1 + n_2 + n_3}, \quad \hat{p}_2 = \frac{n_2}{n_1 + n_2 + n_3} \text{ and } \hat{p}_3 = \frac{n_3}{n_1 + n_2 + n_3}$$

As discussed in the case of 1936 American presidential election, the sample selection plays an important role in unbiased estimate of the proportions. The same applies to estimation of any parameter, such as scale, shape, and location parameters of probability distributions. Later in the chapter, we will discuss the method of moments and maximum likelihood estimation which can be used for estimating parameters of probability distribution.

## 4.3 | SAMPLING

The process of identifying a subset from a population of elements (aka observations or cases) is called sampling process or simply sampling. The following steps are used in any sampling process:

- 1. Identification of target population that is important for a given problem under study.** For example, assume that we are interested in studying attrition among young professionals in India. The definition 'young professionals' in India is vague; we need a clear identification of the target population. A better definition of the population in this case would be to study the attrition among IT professionals in the age group 25–35 years in India. It is important to clearly define the target population for correct inference.
- 2. Decide the sampling frame.** Sampling frame defines the source (or method/procedure) used for identifying the elements of the target population. Choice of sampling frame is important for accuracy of the study. Literary Digest used the telephone directory as one of the sampling frame which turned out to be an incorrect sampling frame. One may use more than one sampling frame (Literary Digest did use more than one sampling frame). Sampling framework will also include features of individual entities. One of the challenges at this stage of sampling process is that in many situations, sampling frame itself may not exist (Kiregyera, 1982). In such cases, the researcher has to define sampling frame using which individual data may be collected. To analyse attrition among IT professionals, sources such as LinkedIn and job portals Naukri and Monster can be used. However, these frames may not have important variables (features) that are required such as information related to salary and other data captured during exit interview. So, ideally to understand the attrition behaviour one has to use the data captured by many human resource departments across multiple companies.
- 3. Determine the sample size:** Determining sample size for data collection is important since collecting data can be expensive and at the same time insufficient sample results in lack of precision in estimation of the parameters. The sample size for analytics projects is determined using factors such as effect size, standard deviation, desired level of confidence, and margin of error. We will discuss the formula for calculating the sample size in section 4.8. An important point to note here is that even in the days of big data in which many business contexts produce huge quantity of data, we still have several scenarios for which sufficient data may not be available (especially when the event itself is rare such as occurrence of Tsunami). Several rules of thumb are often used for determining sample size. For multivariate models such as multiple linear regression, logistic regression, and factor analysis, the thumb rule such as 10 times or 20 times the number of independent variables are used (Norman *et al.*, 2012). That is, if there are 10 variables, then a sample size of 200 in most cases would be acceptable (Norman *et al.*, 2012).
- 4. Sampling method:** Sampling method is the technique used for selecting individual cases in the sample from the target population using the sampling frame. At a higher level, sampling method is classified into two major categories: **probabilistic sampling** and **non-probabilistic sampling**. Probabilistic sampling is further classified as random sampling, stratified sampling, etc. Bootstrap aggregating (also known as Bagging) and Boosting are two popular sampling methods used in machine learning algorithms.

## 4.4 | PROBABILISTIC SAMPLING

In a probability sampling, the individual observations in the sample are selected according to a probability distribution. Assume that the population has a total of  $N$  cases, and we are interested in creating a sample of size  $n$ . There are  ${}^N C_n$  [ $= N! / \{n! \times (N-n)!\}$ ] different ways for creating such a sample. For example, if  $N = 100$  and  $n = 30$ , there will more than  $2.93 \times 10^{25}$  possible samples. Based on how each case in the sample is selected forms the basis of different sampling methods.

### 4.4.1 | Random Sampling

Random sampling is one of the most popular and frequently used sampling methods. Shewhart (1931) defines random sample as a ‘sample drawn under conditions such that the law of large number applies’. That is, in random sampling, every case in the population has equal probability of getting selected in a sample. Random sampling is usually carried out **without replacement**, that is, an observation which is selected in the sample is removed from the population for subsequent selection. However, random samples can also be created **with replacement**, that is, an observation which is selected for inclusion in the sample can again be considered since it is replaced (not removed) in the population.

Selection of cases in a sample can be implemented using several procedures. The easiest one is to label all cases in the population sequentially and generate uniform random numbers (integers) to select the cases from the population. For example, assume that the population has 10 cases (patients and their length of stay measured in days for treatment at a hospital) as shown in Table 4.1.

Now, to generate a sample of size 5 cases, we can generate 5 integer random numbers that follow uniform distribution between 1 and 10 and use those cases in the sample. Uniform random integer numbers between two values can be generated using Microsoft Excel function RANDBETWEEN(lower value, upper value). Table 4.2 shows the random numbers generated using RANDBETWEEN(1, 10) and the corresponding samples (length of stay of patients selected in the sample).

One disadvantage of the above procedure is that the random numbers are likely to repeat, and thus are not ideal for generating samples without replacement. For sampling without replacement, one may generate one case in the sample sequentially, removing and reordering the population at each step. Simple random sampling is used when the population is homogenous.

**TABLE 4.1** Patients and length of stay (LoS) in days

Patient	1	2	3	4	5	6	7	8	9	10
LoS	4	20	12	13	15	17	16	20	9	17

**TABLE 4.2** Random sample of size 5 using uniform random numbers

Random Numbers					Corresponding Sample (LoS value)				
3	4	5	1	8	12	13	15	4	20
1	7	9	1	3	4	16	9	4	12
8	4	7	3	5	20	13	16	12	15



**IMPORTANT**

*Random sampling is ideal when the population is homogeneous. In random sampling, every subject in the population has equal probability of selection in the sample.*

#### 4.4.2 | Stratified Sampling

The population can be divided into mutually exclusive groups using some factor (for example, age, gender, marital status, income, geographical regions, etc.). The groups, thus, formed are called **stratum**. It is important that the groups are mutually exclusive and exhaustive of the population. Examples of few stratified samples are as follows:

1. Amount of time spent by male and female users in sending messages in a day. Here the strata are male and female users.
2. Efficacy of a drug among different age groups. Age group can be classified into categories such as less than 40, between 41 and 60, and over 60 years of age.
3. Performance of children in school and the parents' marital status. Here, marital status can be (a) Single, (b) Married, (d) Divorced. In this case we assume that the parent's marital status may influence children's academic performance.
4. Television rating points for a program across different geographical regions of a country. For India, geographical regions could be different states of the country.

Random sampling method can be used within each stratum to select individual cases to generate samples in each group. Size of the sample in each strata should be proportional to the proportion of the strata in the population. Samples from each stratum can be combined to create the final sample. The following steps are used in creating stratified sample.

1. Identify the factor that can be used for creating strata (for example: factor = Age; Strata 1: age less than 40; Strata 2: age between 41 and 60; and Strata 3: Age more than 60).
2. Calculate the proportion of each stratum in the population (say  $p_1$ ,  $p_2$ , and  $p_3$  be the proportions for three strata identified in step 1).
3. Calculate the sample size (say  $N$ ). The sample size for strata 1, 2, and 3 identified in step 2 are  $p_1 \times N$ ,  $p_2 \times N$ , and  $p_3 \times N$ , respectively.
4. Use random sampling procedure explained in Section 4.4.1 to generate random samples in each strata.
5. Combine samples from each stratum to create the final sample.

Stratified sample is necessary when the population is heterogeneous and creating homogeneous stratum before sampling is recommended for precise estimation of population parameters.

#### 4.4.3 | Cluster Sampling

In cluster sampling, the population is divided into mutually exclusive clusters. For example, assume that a researcher is interested in analysing life of smart phone batteries from a specific manufacturer. The manufacturer may have different models (each model in this case will be a cluster). The clusters are randomly selected and then all units within the selected clusters are included in the sample (or random sampling is used for selecting subjects within the cluster if the cluster size is too large). The following steps are used in cluster sampling:

1. Identify the clusters (example: different models of smart phones sold by a manufacturer, customers from different geographical locations).
2. Use random sampling select the clusters.
3. Select all units in the clusters selected in step 2 and form the sample. If the size is too large, a random sampling within the clusters identified in step 2 may be used for final sample.

Note that stratified sampling and cluster sampling are similar; the major difference is that in a stratified sample, all strata will be represented in the sample, whereas in a cluster sampling, not all clusters will be represented. Cluster sample is used when the clusters are large in number. For example, assume that we are interested in impact of demonetization on Indian industry. There are large number of industrial sectors. Analysing the impact on all clusters will be expensive and time consuming, so in such cases few clusters (such as healthcare and manufacturing) may be used for the study.

#### 4.4.4 | Bootstrap Aggregating (Bagging)

**Bootstrap Aggregating** (known as Bagging) is sampling with replacement used in machine learning algorithms, especially the random forest algorithm (Breiman, 1996). In Bagging, several samples (with replacement) are generated from the population and analytical models are developed using each sample. The size of each sample and the number of samples are determined based on factors such as population size, target accuracy of the model developed using bagging and convergence, etc. Bagging is frequently used in ensemble methods (in which several models are developed and the final prediction is usually based on the majority voting). For example, in random forest, several hundred samples are generated from the population and classification trees are generated using each sample. The final classification of a new case is decided based on the majority voting.

### 4.5 | NON-PROBABILITY SAMPLING

In a non-probability sampling, the selection of sample units from the population does not follow any probability distribution. Sample units are selected based on convenience and/or on voluntary basis. Assume that a data scientist is interested in studying attrition and factors influencing attrition. For this study, he/she may collect data from his friends and colleagues which may not be true representation of the population. Such sampling procedures come under the category of non-probability sampling since

the sample cases are not chosen probabilistically. The accuracy of estimation based on non-probability sampling may be biased.

#### 4.5.1 | Convenience Sampling

Convenience sampling is a non-probability sampling technique in which the sample units are not selected according to a probability distribution. For example, a researcher may collect data from his school or the work place and from his/her friends since the cost of data collection in such cases is minimal. Convenience sampling is not recommended since it is likely to result in biased estimates.

#### 4.5.2 | Voluntary Sampling

Under voluntary sampling the data is collected from people who volunteer for such data collection. For example, customer feedbacks in many contexts fall under this sampling procedure. There could be bias in case of voluntary sampling. Many organizations such as Amazon, Trip Advisor provide customer feedback. Many times the feedback is provided by customers who had bad experience with product/service; many customers who were happy with product/service may not give feedback.

### 4.6 | SAMPLING DISTRIBUTION

In analytics, very often samples are analysed to make inference about the population. Statistic such as mean, proportion, and standard deviation are calculated from a sample and using an appropriate hypothesis test we make inferences about the population parameters. Sample mean is a random variable since different samples drawn from a population are likely to give different sample mean values. Sampling distribution refers to the probability distribution of a statistic such as sample mean and sample standard deviation computed from several random samples of same size. Understanding the sampling distribution is important for hypothesis testing. Test statistic in hypothesis testing is derived based on the knowledge of sampling distribution. For example, consider a population of 10 observations as shown in Table 4.3. We can derive several samples of various sizes from this population of 10 units. For example, there are 100 possible samples of size 2 with replacement and 45 samples of size 2 without replacement. A few samples of size 2 with replacement and corresponding sample means are shown in Table 4.4.

The probability density function of the population is a uniform distribution since each case appears exactly once in the population and the mean value is 27.5. The corresponding probability density function is shown in Figure 4.1.

**TABLE 4.3** Population of 10 observations

S. No.	1	2	3	4	5	6	7	8	9	10
Value	5	10	15	20	25	30	35	40	45	50

**TABLE 4.4** Samples of size 2 and the corresponding mean values

Sample	5, 5	5, 15	10, 5	10, 15	10, 45	20, 5	20, 15	45, 15	50, 20	25, 20
Mean	5	10	7.5	12.5	27.5	12.5	17.5	30	35	22.5

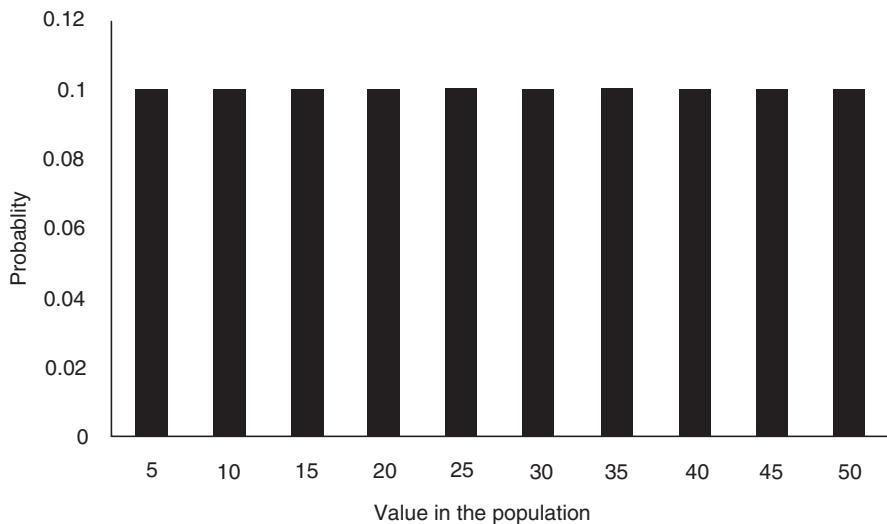


FIGURE 4.1 Probability density function of the population data provided in Table 4.3.

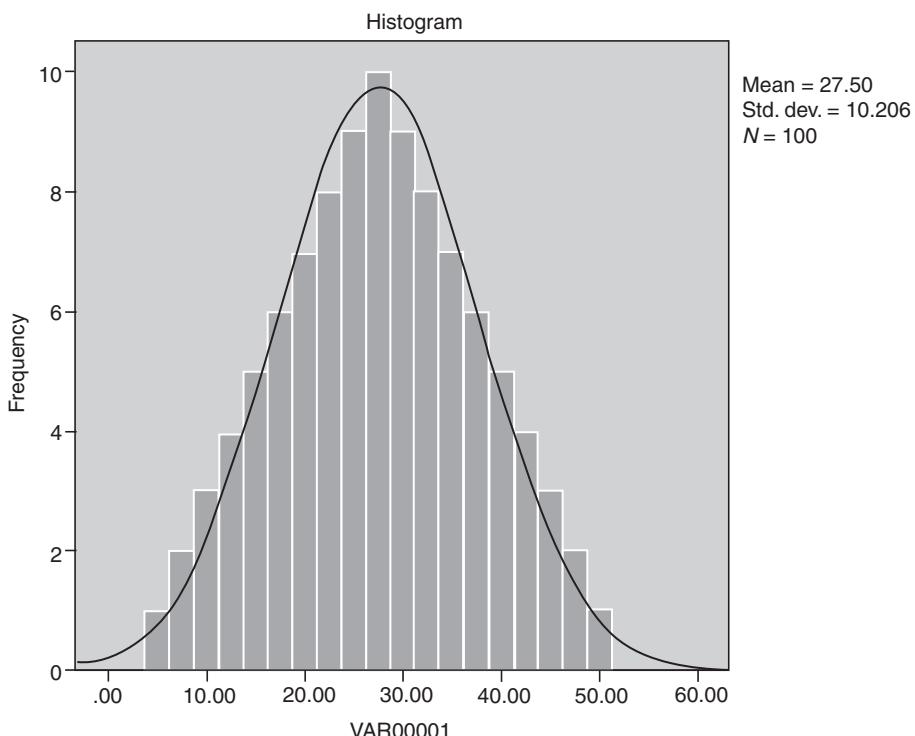


FIGURE 4.2 Histogram of sampling distribution of means.

The histogram of sample mean of all samples of size 2 is shown in Figure 4.2. It is interesting to note that the probability density function of the sampling distribution follows a normal distribution and its mean is 27.5, exactly same as the mean of the population data in Table 4.3.

## 4.7 | CENTRAL LIMIT THEOREM (CLT)

Central limit theorem is one of the most important theorems in statistics due to its applications in testing of hypothesis. Let  $S_1, S_2, \dots, S_k$  be samples of size  $n$  drawn from an independent and identically distributed population with mean  $\mu$  and standard deviation  $\sigma$ . Let  $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_k$  be the sample means (of the samples  $S_1, S_2, \dots, S_k$ ). According to the CLT, the distribution of  $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_k$  follows a normal distribution with mean  $\mu$  and standard deviation  $\sigma / \sqrt{n}$  for large value of  $n$ . That is, the sampling distribution of mean will follow a normal distribution with mean  $\mu$  (same as the mean of the population) and standard deviation  $\sigma / \sqrt{n}$ . The use of the term central limit theorem dates back to the work by George Polya (Fischer, 2010). Several proofs exist for central limit theorem starting from a proof by Laplace in 18<sup>th</sup> century and further works by Poisson and Cauchy (Fischer, 2010).

In simple terms, central limit theorem states that for a large sample drawn from a population with mean  $\mu$  and standard deviation  $\sigma$ , the sampling distribution of mean,  $\bar{X}$ , follows an approximate normal distribution with mean  $\mu$  and standard deviation (standard error)  $\sigma / \sqrt{n}$  irrespective of the distribution of the population (Thomas, 1984).

### Alternative version of CLT can be stated as follows:

Let  $X_1, X_2, \dots, X_n$  be  $n$  random variables that are independent and identically distributed with mean  $\mu$  and standard deviation  $\sigma$ . Then for large  $n$ , mean

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

follows a normal distribution with mean  $\mu$  stand error  $\sigma / \sqrt{n}$ . Independent and identical distribution (IID) implies that the random variables are mutually independent and the random variables follow the same probability distribution.

### Implications of central limit theorem:

1. The variable  $\frac{X - \mu}{\sigma / \sqrt{n}}$  will be a standard normal distribution (mean = 0, standard error = 1).
2. If  $S_n = X_1 + X_2 + \dots + X_n$ , then  $E(S_n) = n\mu$  and Standard error is  $\sigma\sqrt{n}$ . The random variable  $\frac{S_n - n\mu}{\sigma\sqrt{n}}$  is a standard normal variate.
3. Regardless of the population distribution, the sampling distribution of large sample ( $n > 30$ ) will follow the normal distribution with mean same as population mean and standard error  $\sigma / \sqrt{n}$ .

**IMPORTANT**

*Central limit theorem is the basis for hypothesis tests such as Z-test and t-test. In many cases, we will have access to only a sample and the inference about the population has to be made based on sample statistic.*

**IMPORTANT**

*An important assumption of CLT is that the random variables have to be independent and identically distributed.*

#### 4.7.1 | Central Limit Theorem for Proportions

The central limit theorem for proportion is stated as follows:

If we have a population in which a characteristic (for example, smart phone users) has a proportion of  $p$ , then the sampling distribution of the proportion (that is  $\hat{p}$  calculated from several samples of size  $n$ ) will follow a normal distribution with mean  $p$  and standard deviation  $\sqrt{p(1-p)/n}$ .

Central limit theorem for proportions can be stated as follow:

If  $X_1, X_2, \dots, X_n$  are counts from a Bernoulli trials with probability of success  $p$ ,  $E(X_i) = p$  and  $\text{Var}(X_i) = p \times (1-p)$ , then the sampling distribution of probability of success (say  $\hat{p}$ ) follows an approximate normal distribution with mean  $p$  and standard error  $\sqrt{p(1-p)/n}$ , where  $n$  is the sample size. The variable  $\frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$  converges to a standard normal distribution.

**EXAMPLE 4.1**

It is believed that college students in Bangalore spend on average 80 minutes daily on texting using their mobile phones and the corresponding standard deviation is 25 minutes. Data from a sample of 100 students were collected for calculating the amount of time spent in texting. Calculate the probability that the average time spent by this sample of students will exceed 84 minutes.

**Solution:**

Using the central limit theorem, the mean of the sampling distribution is 80 and the corresponding standard deviation is  $25/\sqrt{100} = 2.5$ . The probability that the sample average is more than 84 minutes is given by

$$P\left(Z > \frac{84 - 80}{2.5}\right) = P(Z > 1.6) = 0.05479$$

**EXAMPLE 4.2**

The value of insurance claims received at an insurance company follows exponential distribution with mean INR 4200. If a sample of 100 claims is taken from the population, calculate the probability that the total claim will exceed INR 5,00,000.

**Solution:**

According to CLT, the summation of random variables follows a normal distribution with mean  $n\mu$  and standard error  $\sigma/\sqrt{n}$ . Note that for an exponential distribution mean and standard deviation are same

The probability that the total claim will exceed INR 5,00,000 is

$$P\left(Z > \frac{5,00,000 - n\mu}{\sigma/\sqrt{n}}\right)$$

In this case  $n = 100$ ,  $\mu = \sigma = 4200$ , and  $Z$  is the standard normal variate. So

$$P(Z > 5,00,000) = P\left(Z > \frac{5,00,000 - 100 \times 4200}{4200 \times \sqrt{100}}\right) = P(Z > 1.90476) = 0.02841$$

That is, there is 2.8% chance that the total claim will exceed INR 5,00,000.

## 4.8 | SAMPLE SIZE ESTIMATION FOR MEAN OF THE POPULATION

From the central limit theorem, we know that the sampling distribution of mean follows a normal distribution with mean  $\mu$  and standard deviation  $\sigma/\sqrt{n}$ . Then the standard normal variate of the sampling distribution of mean is given by

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad (4.1)$$

Note that the difference between the sample mean and the population mean,  $\bar{X} - \mu$ , is error in estimation of the population mean. Equation (4.1) can be written as

$$n = \left[ \frac{Z_{\alpha/2} \times \sigma}{D} \right]^2 \quad (4.2)$$

where  $Z_{\alpha/2}$  known as the critical value is the value of  $Z$  for which the area under standard normal distribution is  $\alpha/2$  (that is  $F(Z) = \alpha/2$ ) or  $(1 - \alpha)$  is the desired confidence level in estimating the population mean and  $D = \bar{X} - \mu$  is the error in estimating the population mean. For example, if  $\alpha = 0.05$ , then we have 95% confidence that the error is less than  $D$  when the sample size  $n$  is as given by Eq. (4.2). For  $\alpha = 0.05$ , the value of  $Z_{\alpha/2} = Z_{0.025} = -1.96$  or  $|Z_{\alpha/2}| = 1.96$ . Note that the formula can be used only when the standard deviation is known.

The sample size of estimation of population proportion is given by

$$n = \left( \frac{Z_{\alpha/2}}{D} \right)^2 \times \hat{p} \times (1 - \hat{p}) \quad (4.3)$$

#### EXAMPLE 4.3

A hospital is interested in estimating the average time it takes to discharge a patient after the clearance (discharge note) by the doctor. Calculate the required sample size at a confidence of 95% and maximum error in estimation of 5 minutes. Assume that the population standard deviation is 30 minutes.

**Solution:**

We know that  $D = 5$ ,  $\sigma = 30$ ,  $\alpha = 0.05$ , and  $|Z_{\alpha/2}| = 1.96$  for  $\alpha = 0.05$ . Using Eq. (4.2), we get

$$n = \left[ \frac{Z_{\alpha/2} \times \sigma}{D} \right]^2 = \left[ \frac{1.96 \times 30}{5} \right]^2 \approx 138$$

## 4.9 | ESTIMATION OF POPULATION PARAMETERS

Estimation is a process used for making inferences about population parameters based on samples. For example, we may like to estimate the population parameters such as mean and standard deviation and probability distribution parameters such as scale, shape, and location parameters. The following are the two types of estimates:

- Point Estimate:** Point estimate of a population parameter is the single value (or specific value) calculated from sample (thus called statistic). Sample mean and variance are estimates of population mean and variance. Similarly, sample proportion is an estimate of population proportion.
- Interval Estimate:** Instead of a specific value of the parameter, in an interval estimate the parameter is said to lie in an interval (say between points  $a$  and  $b$ ) with certain probability (or confidence).

The quality of estimates are measured using the following three criteria:

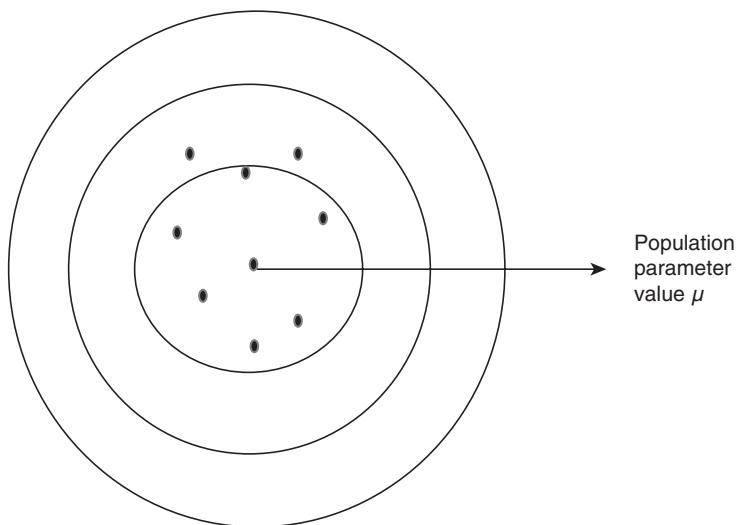
- Unbiased Estimate:** Unbiased estimate of a population parameter is an estimator whose expected value is equal to the population parameter. Let  $\bar{X}$  be an estimate of the population mean  $\mu$ . If  $\bar{X}$  is an unbiased estimate of  $\mu$ , then

$$E(\bar{X}) = \mu \quad \text{or} \quad E(\bar{X} - \mu) = 0$$

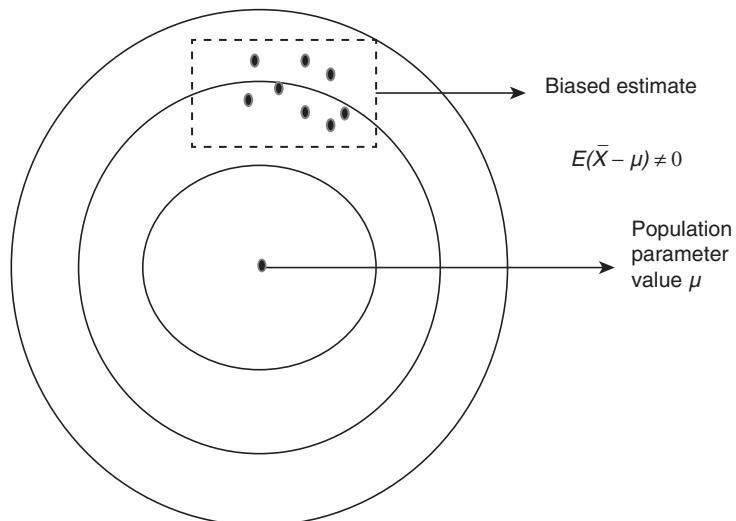
Unbiased estimate is an important requirement of estimation. Figure 4.3 shows the visualization of unbiased estimate and Figure 4.4 shows the visualization of biased estimate.

2. **Consistency:** An estimator of population parameter (say  $\bar{X}$ ) is said to be consistent if it converges to the true value of the parameter ( $\mu$ ) as the size of the sample increases. That is, a consistent estimator implies

$$\lim_{n \rightarrow \infty} \bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n) = \mu \quad (4.4)$$



**FIGURE 4.3** Unbiased estimate (the estimates are randomly scattered around the actual value).



**FIGURE 4.4** Biased estimate (all estimated values are to the one side of the actual value).

3. **Efficiency:** An efficient estimator implies that resulting estimate of the population parameter has the minimum variance.

The estimation of parameters is usually carried out using the following approaches:

1. Method of Moments
2. Maximum Likelihood Estimate (MLE)
3. Bayesian Estimation

We will be discussing method of moments and MLE in the following sections.

## 4.10 | METHOD OF MOMENTS

Moments are measures used in statistics. According to method of moments, a theoretical curve  $Y = f(X, c_1, c_2, \dots, c_n)$ , where  $c_1, c_2, \dots$  are the model parameters, can be fitted given a set of observations by equating the area and first  $(n - 1)$  moments of the observations (Schultz, 1925). The  $n^{\text{th}}$  order moment,  $E(X^n)$ , is given by

$$E(X^n) = \sum_i x_i^n \times p(x_i) \quad \text{when } X \text{ is discrete} \quad (4.5)$$

In Eq. (4.5),  $p(x_i)$  is the probability mass function.

For a continuous random variable, the  $n^{\text{th}}$  moment is given by

$$E(X^n) = \int_{-\infty}^{+\infty} x^n f(x) dx \quad (4.6)$$

In Eq. (4.6),  $f(x)$  is the probability density function.

Central moments are moments about the mean,  $\mu$ , and are given by

$$E(X - \mu)^n = \sum_i (x_i - \mu)^n \times p(x_i) \quad \text{when } X \text{ is discrete} \quad (4.7)$$

For a continuous random variable, the  $n^{\text{th}}$  central moment is given by

$$E(X - \mu)^n = \int_{-\infty}^{+\infty} (x - \mu)^n f(x) dx \quad (4.8)$$

The moments can be connected to various measures of population. The zero-order moment is the total probability.

$$E(X^0) = \int_{-\infty}^{+\infty} x^0 f(x) dx = 1 \quad (4.9)$$

Similarly, the first-order moment is the mean:

$$E(X^1) = \int_{-\infty}^{+\infty} x f(x) dx \quad (4.10)$$

Second-order moment about the mean is the variance:

$$E(X - \mu)^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx \quad (4.11)$$

Thus, moments can be used for estimating the population parameters.

## 4.11 | ESTIMATION OF PARAMETERS USING METHOD OF MOMENTS

Consider a uniform distribution between points  $a$  and  $b$  with probability density function  $[1/(b-a)]$ . We can use the method of moments to estimate the mean and standard deviation as shown below:

$$E(X) = \int_a^b x \times \frac{1}{b-a} dx = \frac{1}{b-a} \int_a^b x dx = \frac{1}{b-a} \left[ \frac{b^2}{2} - \frac{a^2}{2} \right] = \frac{a+b}{2} \quad (4.12)$$

The estimate of variance is

$$E(X - \mu)^2 = \int_a^b [x - (a+b)/2]^2 \times \frac{1}{b-a} dx = \frac{1}{b-a} \times \frac{1}{3} \times \left[ \left( b - \frac{a+b}{2} \right)^3 - \left( a - \frac{a+b}{2} \right)^3 \right] = \frac{1}{12}(b-a)^2 \quad (4.13)$$

### EXAMPLE 4.4

Estimate the expected value of Poisson and exponential random variable using method of moments.

#### Solution:

The probability density function of Poisson distribution is given by

$$f(x) = \frac{e^{-\lambda} \times \lambda^x}{x!}$$

The expected value is given by

$$E(X) = \sum_{i=0}^{\infty} i \times \frac{e^{-\lambda} \times \lambda^i}{i!} = \sum_{i=1}^{\infty} i \times \frac{e^{-\lambda} \times \lambda^i}{i!} = \lambda e^{-\lambda} \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!}$$

Now

$$\sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} = \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = e^{\lambda}$$

$$\text{Thus, } E(X) = \lambda e^{-\lambda} e^{\lambda} = \lambda$$

The probability density of exponential distribution is given by

$$f(x) = \lambda e^{-\lambda x}$$

The expected value is given by

$$E(X) = \int_0^{\infty} x \times \lambda e^{-\lambda x} dx$$

We have to solve the above integration by parts. Let  $u = x$  and  $dv = \lambda e^{-\lambda x} dx$ . Then  $du = dx$  and  $v = -e^{-\lambda x}$ . Integrating by parts we get

$$\int_0^{\infty} x \times \lambda e^{-\lambda x} dx = [uv]_0^{\infty} - \int_0^{\infty} v du = 0 - \left[ \frac{e^{-\lambda x}}{\lambda} \right]_0^{\infty} = \frac{1}{\lambda}$$

Method of moments is simple, but estimates obtained using methods of moments may not be optimal.

## 4.12 | ESTIMATION OF PARAMETERS USING MAXIMUM LIKELIHOOD ESTIMATION

One of the frequently used methods for estimation of parameters of probability distribution is called maximum likelihood estimation (MLE). The main advantages of MLE are that it is mathematically rigorous and less susceptible to individual values as every data in the sample has equal weight in calculation of the estimates of the parameters. The method is very robust and thus can be used for any distribution. Pearson (1936) demonstrated that MLE estimate provided better estimates than method of moments using few examples. The method starts with a belief about the probability distribution of the data and estimates the parameter that maximizes the likelihood function of observing the data given the distribution. MLE has the following steps:

1. Start with a belief about the population (say exponential distribution).
2. Derive the likelihood function that estimates probability of observing the data using the belief in step 1.
3. Take natural logarithmic transformation of the likelihood function (Log likelihood function). Log likelihood function is used to simplify the computation.
4. Estimate the parameters that maximize the log likelihood function derived in step 3.

### 4.12.1 | Estimation of Binomial Distribution Parameter

Consider a binomial distribution with  $n$  Bernoulli trials and each with probability of success  $p$ . The objective is to estimate the probability  $p$ . The probability density function of binomial distribution is given by

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad (4.14)$$

where

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Let  $x_1, x_2, \dots, x_m$  be the number of success obtained out of  $n$  successive trials repeated  $m$  times. The corresponding joint probability is the likelihood of observing  $x_1, x_2, \dots, x_m$  successes out of  $n$  trials repeated  $m$  times and is given by

$$L(x_1, x_2, \dots, x_m | p, n) = \prod_{i=1}^m f(x_i) = \prod_{i=1}^m \binom{n}{x_i} p^{x_i} (1-p)^{n-x_i} \quad (4.15)$$

The log likelihood function is given by

$$LL(x_1, x_2, \dots, x_m | p, n) = \sum_{i=1}^m \ln(n!) - \sum_{i=1}^m \ln(x_i!) - \sum_{i=1}^m \ln(n-x_i!) + \sum_{i=1}^m x_i \ln(p) + \sum_{i=1}^m (n-x_i) \ln(1-p) \quad (4.16)$$

Taking derivative and setting it to zero, we get

$$\frac{dLL(x_1, x_2, \dots, x_m | p, n)}{dp} = \frac{1}{p} \sum_{i=1}^m x_i - \frac{1}{1-p} \sum_{i=1}^m (n-x_i) = 0 \quad (4.17)$$

That is

$$(1-p) \sum_{i=1}^m x_i - p \sum_{i=1}^m (n-x_i) = 0 \Rightarrow \sum_{i=1}^m x_i - p \sum_{i=1}^m x_i - p \times m \times n + p \sum_{i=1}^m x_i = 0$$

Thus, the estimate  $\hat{p}$  is given by

$$\hat{p} = \frac{\sum_{i=1}^m x_i}{m \times n} \quad (4.18)$$

That is, the estimate  $\hat{p}$  is average of proportions.

#### EXAMPLE 4.5

A talent acquisition company is interested in estimating the probability of successful recruitment of top executives for their clients. Table 4.5 shows the number of successful recruits out of 10 persons interviewed during the past 8 recruitment cycles. Estimate the probability of success  $p$  using the maximum likelihood estimation.

**TABLE 4.5** Data on successful recruits

Recruitment cycle number	1	2	3	4	5	6	7	8
Number of people recruited	4	2	5	4	2	1	5	3

**Solution:**

The estimate of  $p$  is given by

$$\hat{p} = \frac{\sum_{i=1}^8 x_i}{m \times n} = \frac{26}{8 \times 10} = 0.325$$

#### 4.12.2 | Estimation of Scale Parameter of Exponential Distribution

Assume that a data set  $\{X_1, X_2, \dots, X_n\}$  follows an exponential distribution with scale parameter  $\lambda$ . The objective of MLE is to estimate the value of  $\lambda$  that will maximize the likelihood of the data  $\{X_1, X_2, \dots, X_n\}$ . The likelihood of observing the data  $\{X_1, X_2, \dots, X_n\}$  from an exponential distribution is given by the joint probability given in Eq. (4.19):

$$L(\lambda) = f(X_1, X_2, \dots, X_n) = \lambda e^{-\lambda X_1} \times \lambda e^{-\lambda X_2} \times \dots \times \lambda e^{-\lambda X_n} = \lambda^n e^{-\lambda \sum_{i=1}^n X_i} \quad (4.19)$$

where  $L(\lambda)$  is the likelihood function, which is same as the joint probability of observing the data  $\{X_1, X_2, \dots, X_n\}$  that follows an exponential distribution. In Eq. (4.19), we assume that the events  $X_1, X_2$ , etc. are independent. The objective of MLE is to find the value of  $\lambda$  that will maximize the likelihood function, that is

$$\underset{\lambda}{\text{Max}} \left( L(\lambda) = \lambda^n \times e^{-\lambda \sum_{i=1}^n X_i} \right) \quad (4.20)$$

To find the optimal value of  $\lambda$ , we have to take the derivative of the likelihood function in Eq. (4.20) and equate that to zero. However, the derivative of Eq. (4.20) is mathematically intractable, thus we take log likelihood function instead of likelihood function defined in Eq. (4.21). The log likelihood function is given by

$$LL(\lambda) = n \ln(\lambda) - \lambda \times \sum_{i=1}^n X_i \quad (4.21)$$

The derivative of Eq. (4.21) with respect to  $\lambda$  is given by

$$\frac{dLL(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^n X_i \quad (4.22)$$

Equating Eq. (4.22) to zero and rearranging, we get

$$\lambda^* = \frac{n}{\sum_{i=1}^n X_i} = \frac{1}{\frac{1}{n} \sum_{i=1}^n X_i} = \frac{1}{\bar{X}} \quad (4.23)$$

where  $\bar{X}$  is the mean value of the observed data.

#### EXAMPLE 4.6

Time to failure of an electronic component is assumed to follow an exponential distribution. Data of 20 failures measured in days are given in Table 4.6. Estimate the time between failure and the failure rate.

**TABLE 4.6** Failure times of 20 electronic components

40	72	56	95	32	12	64	120	145	89
26	37	69	78	98	44	7	21	76	102

**Solution:**

Making the assumption that these times are exponentially distributed, we can find the MLE of the parameter as

$$\lambda = \frac{1}{\bar{X}} = \frac{1}{\frac{1}{20}(40+72+\dots+102)} = \frac{1}{64.15} = 0.01558$$

The estimate of the mean time between failure is 64.15 days and the corresponding failure rate ( $\lambda$ ) is 0.01558.

#### 4.12.3 | MLE of Normal Distribution Parameters

The probability density function for the normal distribution is given by

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The likelihood function of the normal distribution given the data  $\{X_1, X_2, \dots, X_n\}$  is

$$L(X_1, X_2, \dots, X_n; \mu, \sigma) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \quad (4.24)$$

Log likelihood function is given by

$$LL(X_1, X_2, \dots, X_n; \mu, \sigma) = -\sum_{i=1}^n \left\{ \frac{1}{2} \ln(2\pi) + \frac{1}{2} \ln(\sigma^2) + \frac{(X_i - \mu)^2}{2\sigma^2} \right\} \quad (4.25)$$

The maxima of the log likelihood function occurs when

$$\frac{\partial LL}{\partial \mu} = \frac{\partial LL}{\partial \sigma} = 0$$

$$\frac{\partial LL}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu) = 0$$

which can be reduced to

$$\mu = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X} \quad (4.26)$$

That is, the maximum likelihood estimator of the mean is simply the sample mean

$$\frac{\partial LL}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n (X_i - \mu)^2$$

which can be reduced to

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \quad (4.27)$$

## SUMMARY

1. Sampling is a process of creating a subset from the population since collecting the data from the entire population is either expensive or impossible.
2. Sampling process starts by identifying the target population, identifying sampling frame, calculating the sample size, and choosing the method of sampling.
3. Sampling frame which identifies the source of data is important for correct inference about the population. An incorrect sampling frame can result in incorrect inference about the population as demonstrated in the example of Literary Digest.
4. Random sampling, stratified sampling, cluster sampling, and convenient sampling are few frequently used sampling techniques.
5. In a random sampling, every case in the population has equal probability of being selected in the sample. Random sampling is one of the most popular sampling techniques.
6. According to the central limit theorem, sampling distribution of mean and proportion for a large sample follows a normal distribution.
7. Central limit theorem forms the basis of many hypothesis tests and test statistic are derived based on CLT.
8. The estimation of various parameters of probability distributions can be derived using method of moments or using method of maximum likelihood estimation.

## MULTIPLE CHOICE QUESTIONS

1. Estimated value of parameter from a sample is called
  - (a) Sample statistic
  - (b) Population parameter
  - (c) Unbiased estimator
  - (d) Sampling
2. Sampling frame is used for
  - (a) Calculating the sample size
  - (b) Identifying the source of data
  - (c) Estimating population parameter
  - (d) Estimating statistic



**EXERCISES**

- On 8 November 2016, Indian government demonetized 500 and 1000 rupees notes and allowed the citizens to deposit the old notes in the banks. The average amount deposited by customers at a bank in Bannerghatta road is 17500 and the corresponding standard deviation is 4500. If 156 customers deposited money on 11 November 2016.
  - Calculate the probability that the total deposits exceed INR 3 million.
  - What is the probability that the total deposited amount is between INR 2.5 million and INR 3.5 million?
- Average time to churn of a telecom customer is estimated as 300 days from a sample of 10000 customers. What is the probability of observing this sample mean of at least 300 days from a population in which the mean time to churn is 280 days and standard deviation is 72 days?
- The proportion of defaults in mortgage loan is estimated as 8% in the population. In a sample of 1000 mortgage loans, what is the probability that the proportion of defaults will exceed 10%?
- The cost to company (CTC) of 50 IT professionals measured in lakhs of rupees is shown in Table 4.7.

**TABLE 4.7** CTC (in lakhs of rupees)

21.38	12.24	29.06	12.37	8.48	18.76	23.8	28.48	9.56	35.94
28.76	30.76	37.67	34.15	32.53	26.64	24.25	39.66	8.98	26.17
40.54	27.66	18.83	12.87	22.12	28.07	27.15	12.06	5.66	8.44
4.85	11.72	15.18	6.44	28.94	17.71	31.5	26.91	33.93	14.5
38.14	30.87	27.29	6.77	18.43	28.9	22.33	31.41	37.03	32.6

- Draw a histogram. Comment on the distribution of CTC using skewness and kurtosis.
  - Generate 500 random samples of size 10 and plot the histogram of sampling distribution.
  - What is the mean and standard deviation of the sampling distribution obtained in (b)? How far is this mean from the mean CTC of values provided in Table 4.7?
- The amount of time that a vehicle has to wait at a traffic signal in the city of Bangalore is uniformly distributed between 3 and 16 minutes. Use method of moments to estimate the average waiting time and standard deviation of waiting time.
  - Use maximum likelihood estimate to find the scale and shape parameters of a two-parameter Weibull distribution with probability density function

$$f(x) = \frac{\beta}{\eta} \left( \frac{x}{\eta} \right)^{\beta-1} e^{-\left(\frac{x}{\eta}\right)^\beta}$$

where  $\eta$  is the scale parameter and  $\beta$  is the shape parameter.

- Call duration of calls made by customers of a telephone company follows an exponential distribution with mean 320 seconds and standard deviation 80 seconds. Calculate the probability that the average call duration of a random sample of 250 calls will exceed 300 seconds.
- Waiting time at a bank follows a normal distribution with mean 16 minutes and standard deviation 4 minutes. Calculate the sample size required to estimate the mean at a confidence of 95% and maximum error in estimation of 2 minutes.
- According to the government sources, 1% of 1000 rupees currency notes are counterfeit notes. If one would like to estimate the percentage of counterfeit notes in circulation within an error range of 0.001, calculate the sample size required at  $\alpha = 0.01$ .

10. Table 4.8 shows time to failure of air conditioners (ACs) sold by a company measured in days since sale. The time to failure is assumed to follow an exponential distribution. Calculate the mean time between failure ( $1/\lambda$ ) using maximum likelihood estimate. If the warranty period is 365 days, calculate the proportion of ACs likely to fail during the warranty period.

**TABLE 4.8** Time to failure (measured in days) of air conditioners

86	554	318	366	1180	175	74	276	653	438
284	161	32	342	470	701	314	989	586	3151
295	3999	1790	116	272	176	80	215	1770	733
1751	1809	142	888	200	501	237	304	1563	252
106	372	1097	133	145	69	201	3070	957	111

## REFERENCES

1. Brieman L (1996), “Bagging Predictors”, *Machine Learning*, **24**, 123–140.
2. Cahalan D (1989), “The Digest Poll Rides Again”, *The Public Opinion Quarterly*, **53**(1), 129–133.
3. Fischer H (2010), “A History of the Central Limit Theorem”, Springer, Germany.
4. Kiregyera B (1982), “On Sampling Frames in African Censuses and Surveys”, *Journal of Royal Statistical Society, Series D*, **31**(2), 153–167.
5. Mascarenhas A (2016), “World Health Day: India Among Top 3 Countries with high Diabetic Population”, *The Indian Express*, April 7 2016, available at <http://indianexpress.com/article/lifestyle/health/diabetes-cases-422-mn-worldwide-india-no-2-who-lancet-world-health-day/> accessed on 28 April 2017.
6. Moxley N (2012), “Flashback: The Ultimate crying game... when Manchester city and Stoke were relegated”, Mail Online, 24 March 2012.
7. Norman G, Monteiro S and Salama S, (2012), “Sample Size Calculations: Should the Emperor’s Clothes be off the Peg or made to Measure?”, *British Medical Journal*, **345**(7874), 19–21.
8. Pearson K (1936), “Method of Moments and Method of Maximum Likelihood”, *Biometrika*, **28**(1/2), 34–59.
9. Schultz H (1925), “An Extension of the Method of Moments”, *Journal of American Statistical Association*, **20**(150), 242–244.
10. Shewhart W A (1931), “Random Sampling”, *The American Mathematical Monthly*, **38**(5), 245–270.
11. Squire P (1988), “Why the 1936 Literary Digest Poll Failed”, *The Public Opinion Quarterly*, **52**(1), 125–133.
12. Summers G F (1969), “Towards a Paradigm for Respondent Bias in Survey Research”, *The Sociological Quarterly*, **10**(1), 113–121.
13. Szczepanik N (2016), “Pulp Football: An Amazing Anthology of Real Football Stories You Simply Couldn’t Make Up”, Pitch Publishing, Durlington.
14. Thomas D A (1984), “Understanding the Central Limit Theorem”, *The Mathematics Teacher*, **77**(7), 542–543.

# Confidence Intervals

5

“Confidence comes not from always being right but from not fearing to be wrong.”

— Peter McIntyre

## LEARNING OBJECTIVES

- LO 5-1** Learn the difference between point estimate and interval estimate. Understand the need for interval estimate.
- LO 5-2** Understand the concept of confidence interval and confidence level.
- LO 5-3** Learn to calculate confidence interval for population mean when population standard deviation is either known or unknown.
- LO 5-4** Understand confidence interval for population proportion and variance.
- LO 5-5** Gain insights from confidence interval and confidence level.

## CONFIDENCE INTERVALS

When there is an uncertainty around measuring the value of an important population parameter, it is advisable to find the range in which the value of the parameter is likely to fall rather than predicting a single estimate (point estimate). Confidence interval is the range in which the value of a population parameter is likely to lie with certain probability. Confidence interval provides additional information about the population parameter that will be useful in decision making.

IMPORTANT

*The objective of confidence interval is to provide both location and precision of population parameters.*

### 5.1 | INTRODUCTION TO CONFIDENCE INTERVAL

We estimate population parameters such as mean, proportion, standard deviation, scale, shape, and location parameters of the probability distribution from a sample using techniques such as method of moments and maximum likelihood estimation (MLE). Point estimate obtained through techniques such as methods of moments and MLE is a unique value. The quality of estimated parameter values is

measured using factors such as biasness, consistency, and efficiency (discussed in Chapter 4). The accuracy of the point estimate of population parameters is very difficult to establish; hence we prefer *interval estimate* over point estimate. An interval estimate is defined as follows:

An interval estimate of a population parameter such as mean and standard deviation is an interval or range of values within which the true parameter value is likely to lie with certain probability.

The interval estimate is stated between two values. For example, confidence interval for population mean may be stated as  $30 \leq \mu \leq 50$  (that is, the population mean lies between values 30 and 50). The interval estimate may or may not contain the true parameter values. Thus, we associate a confidence (probability) with interval estimate that predicts the probability of finding true parameter value in the interval. For example, we may state that there is a 95% confidence that the interval contains the population mean. 95% confidence also implies that there is a 5% chance that the interval may not contain the actual population mean. Depending on the context of the problem, one may increase confidence level to 98% or 99%. Confidence level is defined as follows:

Confidence level, usually written as  $(1 - \alpha)100\%$ , on the interval estimate of a population parameter is the probability that the interval estimate will contain the true population parameter. When  $\alpha = 0.05$ , 95% is the confidence level and 0.95 is the probability that the interval estimate will have the population parameter.

The value of  $\alpha$  is called *significance*. The value of  $\alpha$  signifies that the chance of not observing the true population mean in the interval estimate is 1 out of 20. Alternatively, 95% confidence implies that in 19 out of 20 cases, the true population mean will be within the interval estimate. The confidence interval is defined as follows:

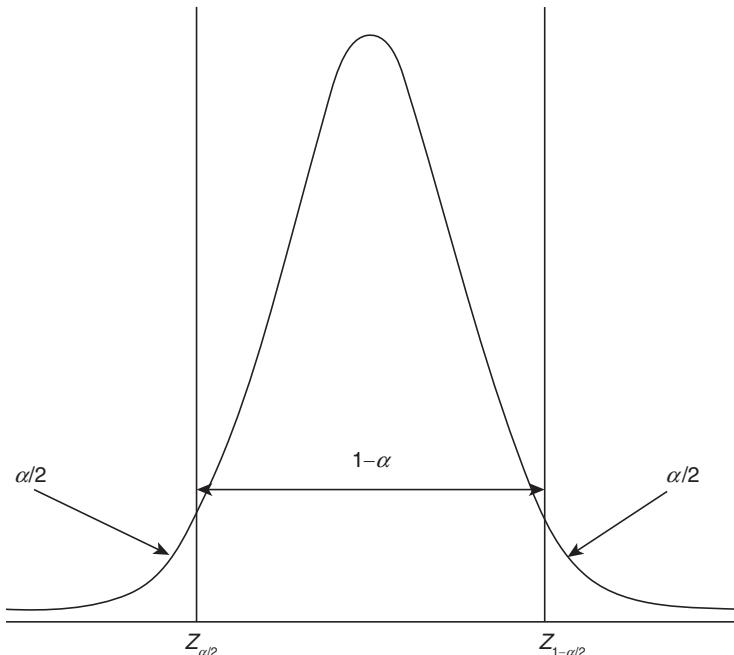
Confidence interval is the interval estimate of the population parameter estimated from a sample using a specified confidence level.

95% is most frequently used confidence level, although 90% and 99% are also used frequently. The choice of  $\alpha$  depends on the context of the problem. When high accuracy for the estimate is required, low value of  $\alpha$  is chosen.

## 5.2 | CONFIDENCE INTERVAL FOR POPULATION MEAN

Let  $X_1, X_2, \dots, X_n$  be the sample means of samples  $S_1, S_2, \dots, S_n$  that are drawn from an independent and identically distributed population with mean  $\mu$  and standard deviation  $\sigma$ . From central limit theorem we know that the sample means  $X_i$  follow a normal distribution with mean  $\mu$  and standard deviation  $\sigma / \sqrt{n}$ . The variable  $Z = \frac{X_i - \mu}{\sigma / \sqrt{n}}$  follows a standard normal variable.

Assume that we are interested in finding  $(1 - \alpha)100\%$  confidence interval for the population mean. We can distribute  $\alpha$  (probability of not observing true population mean in the interval) equally ( $\alpha/2$ ) on either side of the distribution as shown in Figure 5.1.



**FIGURE 5.1** Confidence interval for population mean.

For  $\alpha = 0.05$  or  $\alpha/2 = 0.025$  (that is for 95% confidence level), we can calculate lower and upper values of the confidence interval from the standard normal distribution. Using the standard normal table [one can use  $\text{Normsinv}(\alpha/2)$  or  $\text{Norm.s.inv}(\alpha/2)$  function in Microsoft Excel] we can get the value of  $Z$  for which the area under the normal distribution is less than 0.025. The corresponding value is approximately  $-1.96$ . Since normal distribution is symmetrical distribution, the corresponding value (that is, the value of  $Z$  for which the area is less than 0.975)  $Z_{1-\alpha/2}$  is approximately  $1.96$ . We use the notation  $Z_{\alpha/2}$  to denote the value of standard normal variate for which the area under standard normal distribution is less than or equal to  $\alpha/2$ . That is,  $Z_{0.025} = -1.96$  and  $Z_{0.975} = 1.96$ . (Note that for a normal distribution,  $Z_{1-\alpha/2} = -Z_{\alpha/2}$ ). Using the transformation relationship between standard normal random variable  $Z$  and normal random variable  $X$  ( $X = \mu + Z\sigma$ ), we can write the 95% confidence interval for population mean when population standard deviation is known as:

$$\bar{X} \pm 1.96 \sigma / \sqrt{n} \quad (5.1)$$

$\bar{X}$  is the estimated value of mean from a sample of size  $n$ . In general,  $(1 - \alpha)100\%$  the confidence interval for the population mean when population standard deviation is known can be written as

$$\bar{X} \pm Z_{\alpha/2} \times \sigma / \sqrt{n} \quad (5.2)$$

Equation (5.2) is valid for large sample sizes, irrespective of the distribution of the population. Equation (5.2) is equivalent to

$$P(\bar{X} - Z_{\alpha/2} \times \sigma / \sqrt{n} \leq \mu \leq \bar{X} + Z_{\alpha/2} \times \sigma / \sqrt{n}) = 1 - \alpha \quad (5.3)$$

That is, the probability that the population mean takes a value between  $\bar{X} - Z_{\alpha/2} \times \sigma / \sqrt{n}$  and  $\bar{X} + Z_{\alpha/2} \times \sigma / \sqrt{n}$  is  $1 - \alpha$ . The absolute values of  $Z_{\alpha/2}$  for various values of  $\alpha$  are shown in Table 5.1.

**TABLE 5.1** Value of  $|Z_{\alpha/2}|$  for different values of  $\alpha$

$\alpha$	$ Z_{\alpha/2} $	Confidence interval for population mean when population standard deviation is known
0.1	1.64	$\bar{X} \pm 1.64 \times \sigma / \sqrt{n}$
0.05	1.96	$\bar{X} \pm 1.96 \times \sigma / \sqrt{n}$
0.02	2.33	$\bar{X} \pm 2.33 \times \sigma / \sqrt{n}$
0.01	2.58	$\bar{X} \pm 2.58 \times \sigma / \sqrt{n}$

### EXAMPLE 5.1

A sample of 100 patients was chosen to estimate the length of stay (LoS) at a hospital. The sample mean was 4.5 days and the population standard deviation was known to be 1.2 days.

- (a) Calculate the 95% confidence interval for the population mean.
- (b) What is the probability that the population mean is greater than 4.73 days?

#### Solution:

- (a) **95% confidence interval for population mean:** We know that  $\bar{X} = 4.5$  and  $\sigma = 1.2$  and thus  $\sigma / \sqrt{n} = 1.2 / \sqrt{100} = 0.12$ .

The 95% confidence interval is given by

$$\begin{aligned} (\bar{X} - Z_{\alpha/2} \times \sigma / \sqrt{n}, \bar{X} + Z_{\alpha/2} \times \sigma / \sqrt{n}) &= (4.5 - 1.96 \times 0.12, 4.5 + 1.96 \times 0.12) \\ &= (4.2648, 4.7352) \end{aligned}$$

The Excel function `CONFIDENCE( $\alpha$ ,  $\sigma$ ,  $n$ )` [or `CONFIDENCE.NORM( $\alpha$ ,  $\sigma$ ,  $n$ )`], where  $\alpha$  is the significance,  $\sigma$  is the population standard deviation, and  $n$  is the sample size, returns the value  $Z_{\alpha/2} \times \sigma / \sqrt{n}$ . For current problem `CONFIDENCE(0.05, 1.2, 100) = 0.235196`. The corresponding confidence interval is

$$(4.5 - 0.235196, 4.5 + 0.235196) = (4.2648, 4.7352)$$

- (b) Note that 4.73 is the upper limit of the 95% confidence interval from part (a), thus the probability that the population mean is greater than 4.73 is approximately 0.025.

**EXAMPLE 5.2**

Amount of time (measured in hours) spent by 20 students on an online course is given in Table 5.2. Assuming that the population of time spent follows a normal distribution and standard deviation is 3.1 hours, calculate the 90% confidence interval for the mean time spent by the students.

**TABLE 5.2** Sample time spent by students on an online course

4.7	9.3	8	7.4	9.2	1.7	7.2	8.6	9	6.9
9.2	11.2	7.6	4.9	5.3	2.8	12.3	10.6	5.7	3.8

**Solution:** The estimate mean from the sample is  $\bar{X} = 7.27$  and the sampling distribution's standard deviation is  $\sigma / \sqrt{n} = 3.1 / \sqrt{20} = 0.6932$ .

The 90% confidence interval is given by

$$(\bar{X} - Z_{\alpha/2} \times \sigma / \sqrt{n}, \bar{X} + Z_{\alpha/2} \times \sigma / \sqrt{n}) = (7.27 - 1.64 \times 0.6932, 7.27 + 1.64 \times 0.6932) \\ = (6.1332, 8.4068)$$

### 5.3 | CONFIDENCE INTERVAL FOR POPULATION PROPORTION

The central limit theorem for population proportion is stated as follows:

If  $X_1, X_2, \dots, X_n$  are from Bernoulli trials with probability of success  $p$ , that is,  $E(X_i) = p$  and  $\text{Var}(X_i) = p \times q$  (where  $q = 1 - p$ ), then the sampling distribution of probability of success (say  $\hat{p}$ ) for a large sample size follows an approximate normal distribution with mean  $p$  and standard error  $\sqrt{pq/n}$ , where  $n$  is the sample size. The variable  $\frac{\hat{p} - p}{\sqrt{pq/n}}$  converges to a standard normal distribution. Note that the standard deviation of

the sampling distribution of proportions depends on the value of  $p$  which is unknown. However, for large sample size, the estimate value  $\hat{p}$  will converge to the actual value  $p$ . As a rule of thumb, we set the value of  $n$  such that  $n \times p \times q \geq 10$  (few authors suggest that  $n \times p \times q$  should be at least 15).

The  $(1 - \alpha)100\%$  confidence interval for population proportion  $p$  is given by

$$\hat{p} - Z_{\alpha/2} \sqrt{\frac{\hat{p} \times \hat{q}}{n}} \leq p \leq \hat{p} + Z_{\alpha/2} \sqrt{\frac{\hat{p} \times \hat{q}}{n}} \quad (5.4)$$

**EXAMPLE 5.3**

A retail store was interested in finding the proportion of customers who pay through cash (as against credit or debit card) for the merchandize they buy at the store. From a sample of 100 customers, it was found that 70 customers paid by cash. Calculate the 95% confidence interval for proportion of customers who pay by cash.

**Solution:** In this case,  $n = 100$ ,  $\hat{p} = 70/100 = 0.7$  and  $\hat{q} = 1 - \hat{p} = 0.3$ . Since  $n \times \hat{p} \times \hat{q} = 100 \times 0.7 \times 0.3 = 21 \geq 10$ , we can use the confidence interval equation provided in Eq. (5.4).

$$\begin{aligned}\hat{p} - Z_{\alpha/2} \sqrt{\frac{\hat{p} \times \hat{q}}{n}} &\leq p \leq \hat{p} + Z_{\alpha/2} \sqrt{\frac{\hat{p} \times \hat{q}}{n}} \\ \Rightarrow 0.7 - 1.96 \sqrt{\frac{0.7 \times 0.3}{100}} &\leq p \leq 0.7 + 1.96 \sqrt{\frac{0.7 \times 0.3}{100}} = 0.6102 \leq p \leq 0.7898\end{aligned}$$

That is, the 95% confidence interval for  $p$  is  $(0.6102, 0.7898)$ . That is, we are 95% confident that the interval  $(0.6102, 0.7898)$  contains the true population proportion of the customers who pay by cash.

## 5.4 | CONFIDENCE INTERVAL FOR POPULATION MEAN WHEN STANDARD DEVIATION IS UNKNOWN

When the standard deviation of the population is unknown then we will not be able to use the formula stated in Eq. (5.2). William Gossett (Student, 1908) proved that if the population follows a normal distribution and the standard deviation is calculated from the sample, then the statistic given in Eq. (5.5) will follow a  $t$ -distribution with  $(n-1)$  degrees of freedom.

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} \quad (5.5)$$

Here  $S$  is the standard deviation estimated from the sample (standard error). The  $t$ -distribution is very similar to standard normal distribution; it has a bell shape and its mean, median, and mode are equal to zero as in the case of standard normal distribution. The major difference between the  $t$ -distribution and the standard normal distribution is that  $t$ -distribution has broad tail compared to standard normal distribution. However, as the degrees of freedom increases the  $t$ -distribution converges to standard normal distribution.

The  $(1 - \alpha)100\%$  confidence interval for mean from a population that follows normal distribution when the population mean is unknown is given by

$$\bar{X} \mp t_{\alpha/2, n-1} \times \frac{S}{\sqrt{n}} \quad (5.6)$$

In Eq. (5.6), the value  $t_{\alpha/2, n-1}$  is the value of  $t$  under  $t$ -distribution for which the cumulative probability  $F(t) = 0.025$  when the degrees of freedom is  $(n-1)$ . Here the degrees of freedom is  $(n-1)$  since standard deviation is estimated from the sample. The absolute values of  $t_{\alpha/2, n-1}$  for different values of  $\alpha$  are shown in Table 5.3 along with corresponding  $Z_{\alpha/2}$  values.

**TABLE 5.3** Values of  $|t_{\alpha/2,n-1}|$  and  $|Z_{\alpha/2}|$  for different degrees of freedom ( $df$ )

$\alpha$	$ t_{\alpha/2,10} $	$ t_{\alpha/2,50} $	$ t_{\alpha/2,500} $	$ Z_{\alpha/2} $
0.1	1.812	1.675	1.647	1.64
0.05	2.228	2.008	1.964	1.96
0.02	2.763	2.403	2.333	2.33
0.01	3.169	2.677	2.585	2.58

It is evident from Table 5.3 that the values of  $t_{\alpha/2,n-1}$  and  $Z_{\alpha/2}$  converge for higher degrees of freedom. In fact, as the sample size nears 100, the  $t$ -distribution gets very close to a normal distribution. The values of  $t_{\alpha/2,n-1}$  can be obtained using the function T.INV( $\alpha/2, n - 1$ ) in Microsoft Excel [another excel function T.INV.2T( $\alpha, n - 1$ ) also returns  $t_{\alpha/2,n-1}$  value]. In Excel, TINV( $\alpha, n - 1$ ) returns critical values for two-tailed test (concept of two tailed test will be discussed in Chapter 6). If we have to calculate one-tailed critical value at significance  $\alpha$ , then the corresponding Excel function is TINV( $2\alpha, n - 1$ ). Note that, different versions of Microsoft Excel has different functions to calculate inverse value of  $t$  distribution.

**EXAMPLE 5.4**

An online grocery store is interested in estimating the basket size (number of items ordered by the customer) of its customer order so that it can optimize its size of crates used for delivering the grocery items. From a sample of 70 customers, the average basket size was estimated as 24 and the standard deviation estimated from the sample was 3.8. Calculate the 95% confidence interval for the basket size of the customer order.

**Solution:** We know that  $n = 70$ ,  $\bar{X} = 24$ ,  $S = 3.8$  and  $t_{0.025, 69} = 1.995$  [using TINV(0.05, 69) in Microsoft Excel].

The confidence interval for size of basket using Eq. (5.6) is given by

$$\bar{X} \pm t_{\alpha/2,n-1} \frac{S}{\sqrt{n}} = 24 \pm 1.995 \frac{3.8}{\sqrt{70}} = 24 \pm 0.9061$$

Thus the 95% confidence interval for the size of the basket is (23.09, 24.91).

## 5.5 | CONFIDENCE INTERVAL FOR POPULATION VARIANCE

Let  $S_1^2, S_2^2, \dots, S_k^2$  be the sample variance estimated from samples of size  $n$  drawn from a normal distribution with variance  $\sigma^2$ . Then the random variable defined by

$$\frac{(n-1) \times S_i^2}{\sigma^2} \quad (5.7)$$

follows a  $\chi^2$ -distribution with  $(n - 1)$  degrees of freedom. Note that Eq. (5.7) is valid only when the samples are drawn from a normal population; it is not valid otherwise. We can use Eq. (5.7) to derive confidence interval for variance when the samples are drawn from a normal distribution. The  $(1 - \alpha)100\%$  confidence interval for variance,  $\sigma^2$ , is given by (Tate and Klett, 1959 and Cohen, 1972)

$$\left[ \frac{(n-1)S^2}{\chi_{\alpha/2,n-1}^2}, \frac{(n-1)S^2}{\chi_{1-\alpha/2,n-1}^2} \right] \quad (5.8)$$

The confidence interval for standard deviation,  $\sigma$ , is given by

$$\left[ \sqrt{\frac{(n-1)S^2}{\chi_{\alpha/2,n-1}^2}}, \sqrt{\frac{(n-1)S^2}{\chi_{1-\alpha/2,n-1}^2}} \right] \quad (5.9)$$

where  $\chi_{\alpha/2,n-1}^2$  is the value of chi-square distribution with  $n - 1$  degrees of freedom where  $\alpha/2$  is the right side area,  $\chi_{1-\alpha/2,n-1}^2$  is the value of chi-square distribution with  $n - 1$  degrees of freedom where  $1 - \alpha/2$  is the right side area.

### EXAMPLE 5.5

Time taken to manufacture an aircraft door is a random variable due to several manual processes and assembly of more than 1000 parts to make the aircraft door. The sources of variability in door assembly include factors such as non-availability of parts, manpower, and machine tools. It is known that the time to assemble a door follows a normal distribution. The variance of the time taken to manufacture the door was estimated to be 324 hours based on a sample of 50 doors. Calculate a 95% confidence interval for the variance in manufacturing aircraft door.

**Solution:** We know that  $n = 50$ ,  $S^2 = 324$ ,  $\chi_{0.025,49}^2 = 70.22$ ,  $\chi_{0.975,49}^2 = 31.55$  [the value of  $\chi^2$  can be calculated using Microsoft Excel function CHIINV( $\alpha/2$ ,  $df$ ) or CHISQ.INV.RT( $1 - \alpha/2$ ,  $df$ )].

The 95% confidence interval for variance is given by

$$\left[ \frac{(n-1)S^2}{\chi_{\alpha/2,n-1}^2}, \frac{(n-1)S^2}{\chi_{1-\alpha/2,n-1}^2} \right] = \left[ \frac{49 \times 324}{70.22}, \frac{49 \times 324}{31.55} \right] = [226.09, 503.20]$$

The 95% confidence interval for standard deviation is [15.04, 22.43].

### SUMMARY

1. The point estimate of population parameters give unique value, however, data scientists would like to know the range of values the population parameter is likely to take. Interval estimates provide better insights about the population parameter.
2. Confidence level  $(1 - \alpha) \times 100\%$  is the probability that the true population parameter value will lie within the confidence interval.
3. The choice of confidence level or significance  $\alpha$  will depend on the context and accuracy required. For higher accuracy, the value of  $\alpha$  will be lower.
4. Confidence intervals are derived using the central limit theorem of sampling distribution.

**MULTIPLE CHOICE QUESTIONS**

1. 95% confidence interval (CI) for mean is predicted with a sample of 50 cases (say sample 1) and another sample of 100 cases (say sample 2) drawn randomly from the same population. Which of the following statements is true?
  - (a) CI of sample 1 will be smaller than CI of sample 2
  - (b) CI of sample 2 will be smaller than CI of sample 1
  - (c) Sample with small standard deviation will have smaller CI
  - (d) Cannot say
2. 90% confidence interval for number of paid news items telecast by English news channels in India in a week is [20, 30]. The sample mean  $\bar{X}$  in this case is
  - (a) 25
  - (b) 20
  - (c) 30
  - (d) lies between 20 and 30
3. Consider confidence intervals for mean calculated at  $(1 - \alpha)100\%$  and  $(1 - 2\alpha)100\%$ . Which of the following statements are true?
  - (a) The length of the confidence interval at  $(1 - 2\alpha)100\%$  will be twice that of  $(1 - \alpha)100\%$ .
  - (b) The length of the confidence interval at  $(1 - 2\alpha)100\%$  will be smaller than that of  $(1 - \alpha)100\%$ .
  - (c) The length of the confidence interval at  $(1 - 2\alpha)100\%$  will be larger than that at  $(1 - \alpha)100\%$ .
  - (d) It will depend on the sample size of the sample.
4. As the size of the sample increases, the confidence interval for mean at all significance levels will
  - (a) increase
  - (b) decrease
  - (c) not change
  - (d) Cannot say
5. An e-tailer (electronic retailer) estimates the proportion of returns (of the items purchased) by the customers as 0.08 (8%). To estimate the CI, the minimum sample size should be
  - (a) 30
  - (b) 136
  - (c) 68
  - (d) 80
6. Confidence interval for variance of a population can be estimated from a sample only when
  - (a) The sample size is at least 30
  - (b) Population mean is known
  - (c) Population follows a normal distribution
  - (d) Population follows a  $\chi^2$  distribution

**EXERCISES**

1. When Indian Government demonetized the high value currencies (INR 500 and INR 1000) in November 2016, many citizens deposited the demonetized currencies at various bank. At a specific bank, based on a sample of 200 customers, the average demonetized amount deposited was INR 15200 and the standard deviation was 3800. Calculate the 99% confidence interval for the average value of demonetized currencies deposited by customers.
2. At an insurance company, based on a sample of 550 customers, the proportion of fraudulent claims was estimated as 0.07 (7%). Calculate the 99% confidence interval for the proportion of fraudulent claims.
3. At the New Market in Kolkata, the amount of time the customers parked their car (in minutes) is given in Table 5.4. Calculate the 95% confidence interval for the average duration that customers park their car at this mall.

**TABLE 5.4** Parking time at a shopping mall for 30 customers

64	96	122	53	166	153	129	124	42	22
236	232	201	191	200	189	211	181	238	238
99	47	417	46	101	373	43	120	211	224

4. The average delay of flights arriving at the Bangalore international airport follows a normal distribution. A sample of 50 flight delays (in minutes) is provided in Table 5.5. In Table 5.5, negative values indicate arrival before scheduled arrival time and 0 indicates arrival on time. Calculate the 99% confidence interval for the population mean and standard deviation.

**TABLE 5.5** Sample delay (in minutes) of arriving flights

10	-16	11	17	14	-5	3	0	2	3
42	23	30	-3	45	24	30	43	44	-11
6	7	7	-6	7	9	-3	6	5	7
18	-9	20	-9	19	8	16	22	3	23
-18	59	42	-9	-9	42	51	-5	29	31

5. The proportion of undecided voters across 50 districts of a country one-week prior to the election is given in Table 5.6. Calculate the 90% confidence interval for the proportion of undecided voters.

**TABLE 5.6** Proportion of undecided voters one week prior to election

12	16	12	10	14	9	8	13	5	5
19	8	6	11	19	14	10	20	11	10
6	6	5	12	16	9	5	9	17	18
15	17	18	13	18	11	7	20	6	11
23	10	24	6	24	18	7	8	5	15

6. The average television rating points (TRP) for a Hindi television serial estimated on the basis of 50 episodes is 2%, where TRP is the proportion of people watching the serial. Calculate 90% confidence interval for the mean value of population TRP for the serial.
7. Time taken to resolve a customer complaint at a financial services company follows a normal distribution. The variance of resolution time based on 120 complaints was estimated as 25 hours. Calculate the 95% confidence interval for standard deviation of the resolution time.

## REFERENCES

1. Cohen A (1972), "Improved Confidence Interval for the Variance of a Normal Distribution", *Journal of the American Statistical Association*, 67(338), 382–387.
2. Student (1908), "The Probable Error of Means", *Biometrika*, 6, 1–25.
3. Tate R F and Klett G W (1959), "Optimal Confidence Interval for Variance of a Normal Distribution", *Journal of the American Statistical Association*, 54(287), 674–682.

# 6

# Hypothesis Testing

“Beware of the problem of testing too many hypotheses; the more you torture the data, the more likely they are to confess, but confessions obtained under duress may not be admissible in the court of scientific opinion.”

— Stephen M Stigler

## LEARNING OBJECTIVES

- LO 6-1** Understand hypothesis test and its importance in analytics.
- LO 6-2** Learn to setup a hypothesis test, understand the concept of null and alternative hypotheses.
- LO 6-3** Understand the link between central limit theorem and test statistic in one-sample Z-test and *t*-test.
- LO 6-4** Understand the concept of significance ( $\alpha$ ), probability value (*p*-value), Type I and Type II errors.
- LO 6-5** Understand simple one-sample hypothesis test for population mean when population variance is either known or unknown.
- LO 6-6** Learn to conduct a two-sample hypothesis test and its applications in analytics.
- LO 6-7** Understand the role of non-parametric tests such as chi-square test of independence.
- LO 6-8** Learn goodness of fit tests and their application in identifying best probability distribution to describe a data set.

## HYPOTHESIS TESTING

Hypothesis testing is one of the most important concepts in analytics and also a concept which many students of statistics and analytics find it difficult to understand. Hypothesis is a claim made by a person/organization. The claim is usually about population parameters such as mean or proportion and we seek evidence from a sample for the support of the claim (for example, claim could be that the average salary of analytics experts is at least USD 1,00,000). Hypothesis testing is a process used for either rejecting or retaining a null hypothesis.

## IMPORTANT

*The objective of hypothesis testing is to either reject or retain a null hypothesis. In many cases, for example, in regression models, one would like to reject the null hypothesis to establish statistically significant relationship between the dependent and the independent variables. However, in goodness of fit tests, that are used for checking whether the data follows a specific distribution or not, we would like to retain the null hypothesis.*

## 6.1 | INTRODUCTION TO HYPOTHESIS TESTING

### 6.1.1 | Blackout Babies

On 9 November 1965 there was a power failure that resulted in blackout for approximately 12 hours in New York and surrounding areas. Nine months later, in August 1966, New York Times published a series of three articles in which it claimed that the birth rates in August 1966 was higher than normal based on interviews with city doctors (Izenman and Zabell, 1981). The babies were nicknamed ‘blackout babies’. The articles published by the New York Times raised an interesting question on whether power failures result in procreation? Izenman and Zabell (1981) using time series data analysis claimed that there is not enough evidence to suggest that the 1965 power failure resulted in increased birth rate nine months after the blackout. Many claims were made about the impact of power cuts on baby booms and mothers since then (Anon, 2009 and Fetzer *et al.*, 2013).

Hypothesis is a claim or belief, hypothesis testing is a statistical process of either rejecting or retaining a claim or belief or association related to a business context, product, service, processes, etc. Hypothesis testing consists of two complementary statements called **null hypothesis** and **alternative hypothesis**, and only one of them is true. Hypothesis testing is one of the most important concepts in analytics due to its role in inferential statistics. Hypothesis testing is an integral part of many predictive analytics techniques such as multiple linear regression and logistic regression. It plays an important role in providing evidence of an association relationship between an outcome variable and predictor variables.

In business, many claims are made by organizations. Few examples of such claims are listed below:

1. Children who drink the health drink Complan (a health drink owned by the company Heinz in India) are likely to grow taller.
2. If you drink Horlicks, you can grow taller, stronger, and sharper (3 in 1).
3. Using fair and lovely (fair and handsome) cream can make one fair and lovely (fair and handsome).
4. Wearing perfume (such as Axe) will help to attract opposite gender (known as Axe effect).
5. Women use camera phone more than men (Freier, 2016).
6. Beautiful people are likely to have girl child (Miller and Kanazawa, 2007). This is one of my favorite hypotheses since I have a daughter I can claim that I am good looking.
7. Married people are happier than singles (Anon, 2015), especially those who married their best friend (many married people may not agree!).
8. Vegetarians miss few flights (Siegel, 2016).
9. Smokers are better sales people.

There are many such claims and beliefs; many business rules and strategies are generated based on these hypotheses. The question is how can we check whether these are actually true. Hypothesis testing is used for checking the validity of the claim using evidence found in a sample data.

## 6.2 | SETTING UP A HYPOTHESIS TEST

In this section, we will discuss the steps involved in hypothesis testing. Data analysis in general can be classified as **exploratory data analysis** or **confirmatory data analysis**. In exploratory data analysis, the idea is to look for new or previously unknown hypothesis or suggest hypotheses. In the case of confirmatory data analysis, the objective is to test the validity of a hypothesis (confirm whether the hypothesis is true or not) using techniques such as hypothesis testing and regression. According to Tukey (1977), exploratory data analysis is similar to a detective work suggesting hypotheses whereas confirmatory data analysis looks for evidence in support of hypotheses using techniques such as hypothesis testing. The following steps are used in hypothesis testing:

1. Describe the hypothesis in words. Hypothesis is described using a population parameter (such as mean, standard deviation, proportion, etc.) about which a claim (hypothesis) is made. Few sample claims (hypothesis) are:
  - (a) Average time spent by women using social media is more than men.
  - (b) On average women upload more photos in social media than men.
  - (c) Customers with more than one mobile handsets are more likely to churn.
2. Based on the claim made in step 1, define null and alternative hypotheses. Initially we believe that the null hypothesis is true. In general, null hypothesis means that there is no relationship between the two variables under consideration (for example, null hypothesis for the claim ‘women use social media more than men’ will be ‘there is no relationship between gender and the average time spent in social media’). Null and alternative hypotheses are defined using a population parameter.
3. Identify the test statistic to be used for testing the validity of the null hypothesis. Test statistic will enable us to calculate the evidence in support of null hypothesis. The test statistic will depend on the probability distribution of the sampling distribution; for example, if the test is for mean value and the mean is calculated from a large sample and if the population standard deviation is known, then the sampling distribution will be a normal distribution and the test statistic will be a  $Z$ -statistic (standard normal statistic).
4. Decide the criteria for rejection and retention of null hypothesis. This is called **significance value** traditionally denoted by symbol  $\alpha$ . The value of  $\alpha$  will depend on the context and usually 0.1, 0.05, and 0.01 are used. Significance value  $\alpha$  is the Type I error (discussed in Section 6.4).
5. Calculate the  $p$ -value (probability value), which is the conditional probability of observing the test statistic value when the null hypothesis is true. In simple terms,  $p$ -value is the evidence in support of the null hypothesis.
6. Take the decision to reject or retain the null hypothesis based on the  $p$ -value and significance value  $\alpha$ . The null hypothesis is rejected when  $p$ -value is less than  $\alpha$  and the null hypothesis is retained when  $p$ -value is greater than or equal to  $\alpha$ .

### 6.2.1 | Description of Hypothesis

Hypotheses are claims that are usually stated in simple words initially as listed below:

1. Average annual salary of machine learning experts is different for males and females.
2. On an average people with Ph.D. in analytics earn more than people with Ph.D. in engineering.
3. The average box-office collection of comedy genre movies is more than that of action movies.
4. Average life of vegetarians is more than meat eaters.
5. Proportion of married people defaulting on loan repayment is less than proportion of singles defaulting on loan repayment.

### 6.2.2 | Null and Alternative Hypothesis

Null hypothesis, usually denoted as  $H_0$  ( $H$  zero and  $H$  naught), refers to the statement that there is no relationship or no difference between different groups with respect to the value of a population parameter. Null hypothesis is the claim that is assumed to be true initially. That is at the beginning we assume that the null hypothesis is true and try to retain it unless there is strong evidence against null hypothesis.

Alternative hypothesis, usually denoted as  $H_A$  (or  $H_1$ ), is the complement of null hypothesis. Alternative hypothesis is what the researcher believes to be true and would like to reject the null hypothesis.

The null and alternative hypotheses for the sample hypotheses stated in Section 6.2.1 are described in Table 6.1.

**TABLE 6.1** Hypothesis statement to definition of null and alternative hypothesis

S. No.	Hypothesis Description	Null and Alternative Hypothesis
1	Average annual salary of machine learning experts is different for males and females.  (In this case, the null hypothesis is that there is no difference in male and female salary of machine learning experts)	$H_0: \mu_m = \mu_f$ $H_A: \mu_m \neq \mu_f$  $\mu_m$ and $\mu_f$ are average annual salary of male and female machine learning experts, respectively.
2	On average people with Ph.D. in analytics earn more than people with Ph.D. in engineering.	$H_0: \mu_a \leq \mu_e$ $H_A: \mu_a > \mu_e$  $\mu_a$ = Average annual salary of people with Ph.D. in analytics. $\mu_e$ = Average annual salary of people with Ph.D. in engineering. It is essential to have the equal sign in null hypothesis statement.

**IMPORTANT**

*Hypothesis test checks the validity of the null hypothesis based on the evidence from the sample. At the beginning of the test, we assume that the null hypothesis is true. Since the researcher may believe in alternative hypothesis, she/he may like to reject the null hypothesis. However, in many cases (such as goodness of fit tests), we would like to retain or fail to reject the null hypothesis.*

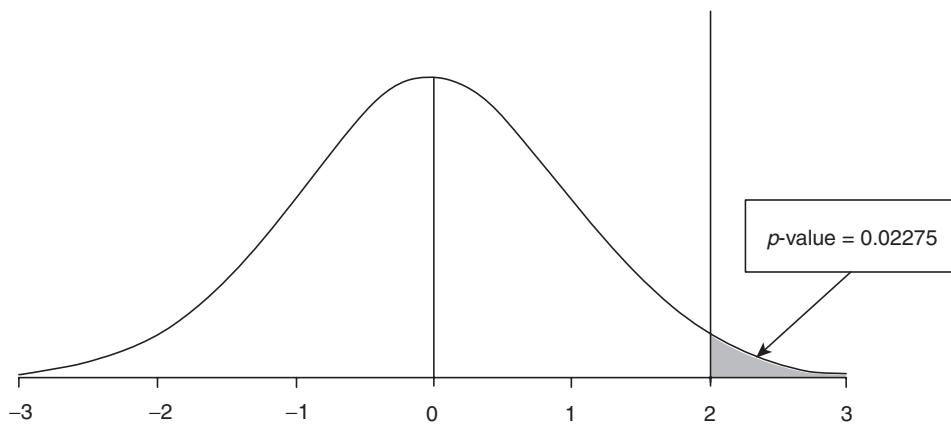
### 6.2.3 | Test Statistic

Test statistic is the standardized difference between the estimated value of the parameter being tested calculated from the sample(s) and the hypothesis value (that is, standardized difference between  $\bar{X}$  and  $\mu$  in the case of testing mean) in order to establish the evidence in support of the null hypothesis. Test statistic is the standardized value used for calculating the  $p$ -value (probability value) in support of null hypothesis. Since test statistic is a standardized value, it measures the standardized distance (measured in terms of number of standard deviations) between the value of the parameter estimated from the sample(s) and the value of the null hypothesis.

The  $p$ -value is the conditional probability of observing the statistic value when the null hypothesis is true. For example, consider the following research hypothesis: Average annual salary of machine learning experts is at least 100,000. The corresponding null hypothesis is  $H_0: \mu_m \leq 100,000$ . Assume that estimated value of the salary from a sample is 1,10,000 (that is  $\bar{X} = 1,10,000$ ) and assume that the standard deviation of population is known and standard error of the sampling distribution is 5000 (that is,  $\sigma / \sqrt{n} = 5000$ , where  $n$  is the sample size using which  $\bar{X} = 1,10,000$  was calculated). The standardized distance between estimated salary from hypothesis salary is  $(1,10,000 - 1,00,000)/5000 = 2$ . That is, the standardized distance between estimated value and the hypothesis value is 2 and we can now find the probability of observing this statistic value from the sample if the null hypothesis is true (that is if  $\mu_m \leq 100,000$ ). A large standardized distance between the estimated value and the hypothesis value will result in a low  $p$ -value. Note that the value 2 is actually the value under a standard normal distribution since it is calculated from  $\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ . Standard normal distribution and the  $p$ -value corresponding to  $Z = 2$  are shown in Figure 6.1.

Probability of observing a value of 2 and higher from a standard normal distribution is 0.02275. That is, if the population mean is 1,00,000 and the standard error of the sampling distribution is 5000 then probability of observing a sample mean greater than or equal to 1,10,000 is 0.02275. The value 0.02275 is the  $p$ -value, which is the evidence in support of the statement in the null hypothesis.

$$p\text{-value} = P(\text{Observing test statistics value} \mid \text{null hypothesis is true}) \quad (6.1)$$



**FIGURE 6.1** Standard normal distribution and  $p$ -value.

**IMPORTANT**

Note that the  $p$ -value is a conditional probability. It is the conditional probability of observing the statistic value given that the null hypothesis is true.  $P$ -value is the evidence in support of null hypothesis.

#### 6.2.4 | Decision Criteria – Significance Value

Primary task in hypothesis testing is to take a decision to either reject or fail to reject (retain) the null hypothesis, thus we need a criteria to take the decision. Significance level, usually denoted by  $\alpha$ , is the criteria used for taking the decision regarding the null hypothesis (reject or retain) based on the calculated  $p$ -value. The significance value  $\alpha$  is the maximum threshold for  $p$ -value. The decision to reject or retain will depend on whether the calculated  $p$ -value crosses the threshold value  $\alpha$  or not. The decision criteria is shown in Table 6.2.

The chosen value of  $\alpha$  may depend on the context of the problem. Usually  $\alpha = 0.05$  is used by researchers (recommended by Fisher, 1956); however, values such as 0.1, 0.02, and 0.01 are also frequently used. The value of  $\alpha$  chosen is very low (0.05) for reason that we start the process of hypothesis testing with an assumption that null hypothesis is true. Unless there is strong evidence against this assumption, we will not reject the null hypothesis. The value of statistic in the sampling distribution for which the probability is  $\alpha$  is called the **critical value**. In a right-tailed test, if the calculated statistic value is greater than the critical value ( $p$ -value will be less than  $\alpha$ -value) then we reject the null hypothesis, whereas, if the statistic value is less than the critical value then we retain the null hypothesis. In case of left-tailed test, if the calculated statistic value is less than the critical value ( $p$ -value will be less than  $\alpha$ -value) then we reject the null hypothesis, whereas, if the statistic value is greater than the critical value then we retain the null hypothesis. The areas beyond the critical values are known as **rejection region**.

**IMPORTANT**

*The significance value  $\alpha$  is the threshold conditional probability of rejecting a null hypothesis when it is true. It is the value of Type I error.*

$$\text{Significance value } \alpha = P(\text{Rejecting a null hypothesis} \mid \text{null hypothesis is true}) \quad (6.2)$$

### 6.3 | ONE-TAILED AND TWO-TAILED TEST

Consider the following three hypotheses:

1. Salary of machine learning experts on average is at least US \$100,000.
2. Average waiting time at the London Heathrow airport security check is less than 30 minutes.
3. Average annual salaries of male and female MBA students are different at the time of graduation.

**TABLE 6.2** Decision making under hypothesis testing

Criteria	Decision
$p\text{-value} < \alpha$	Reject the null hypothesis
$p\text{-value} \geq \alpha$	Retain (or fail to reject) the null hypothesis

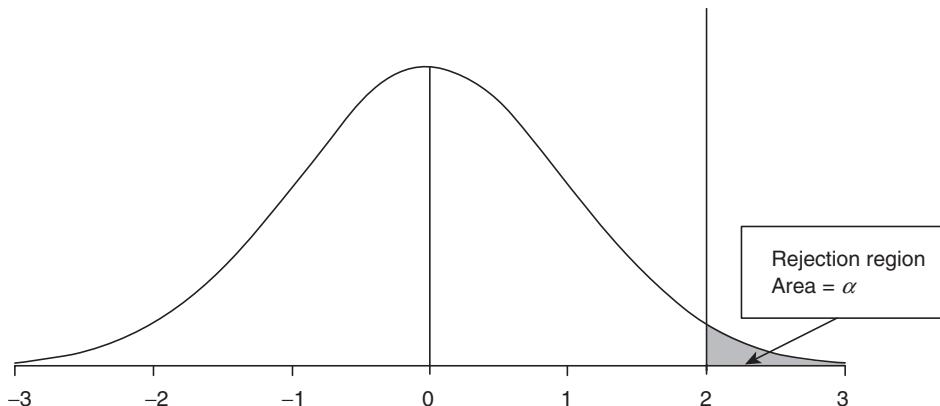
**STATEMENT 1** Salary of machine learning experts on average is at least US \$100,000:

The null and alternative hypotheses in this case are given by

$$H_0: \mu_m \leq 100,000$$

$$H_A: \mu_m > 100,000$$

where  $\mu_m$  is the average annual salary of machine learning experts. Note that the equality symbol is always part of the null hypothesis since we have to measure the difference between estimated value from the sample and the hypothesis value. In this case, reject or retain decision will depend on the direction of deviation of the estimated parameter value from the hypothesis value. Figure 6.2 shows the rejection region on the right side of the distribution. Since the rejection region is only on one side this is a one-tailed test (right tailed test). Specifically, since the alternative hypothesis in this case is  $\mu_m > 100,000$ , this is called right-tailed test.

**FIGURE 6.2** Right-tailed hypothesis test's rejection region.

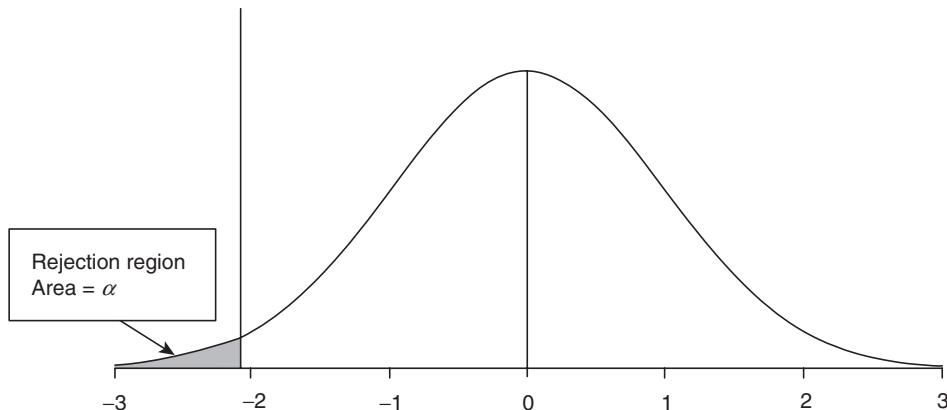
**STATEMENT 2** Average waiting time at the London Heathrow airport security check is less than 30 minutes:

The null and alternative hypotheses in this case are given by

$$H_0: \mu_w \geq 30$$

$$H_A: \mu_w < 30$$

where  $\mu_w$  is the average waiting time at London Heathrow security check. In this case, reject region will be on the left side (known as left-tailed test) of the distribution as shown in Figure 6.3.



**FIGURE 6.3** Rejection region in case of left-sided test.

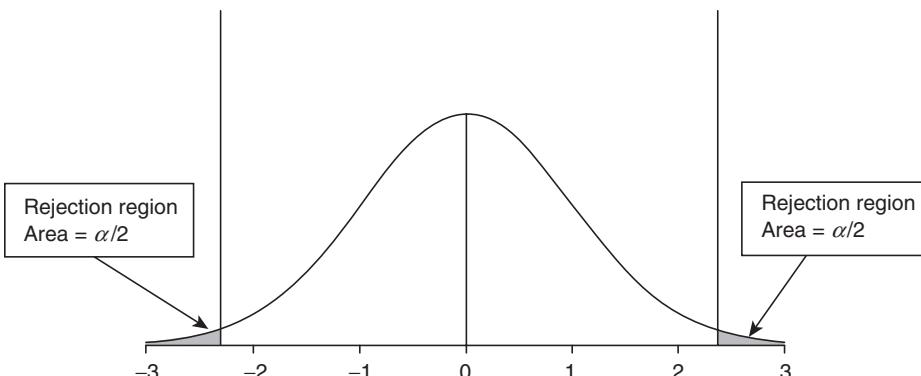
**STATEMENT 3** Average salary of male and female MBA students at graduation is different:

The null and alternative hypotheses in this case are given by

$$H_0: \mu_m = \mu_f$$

$$H_A: \mu_m \neq \mu_f$$

where  $\mu_m$  and  $\mu_f$  are the average salaries of male and female MBA students, respectively, at the time of graduation. In this case, the rejection region will be on either side of the distribution and if the significance level is  $\alpha$  then the rejection region will be  $\alpha/2$  on either side of the distribution. Since the rejection region is on either side of the distribution, it will be a two-tailed test. Figure 6.4 shows the rejection region of a two-tailed test.



**FIGURE 6.4** Rejection region in case of two-tailed test.

## 6.4 | TYPE I ERROR, TYPE II ERROR, AND POWER OF THE HYPOTHESIS TEST

In hypothesis test we end up with the following two decisions:

1. Reject null hypothesis.
2. Fail to reject (or retain) null hypothesis.

Type I and Type II errors are defined as follows:

1. **Type I Error:** Conditional probability of rejecting a null hypothesis when it is true is called Type I Error or False Positive (falsely believing that the claim made in alternative hypothesis is true). The significance value  $\alpha$  is the value of Type I error. Mathematically, Type I error can be defined as follows:

$$\text{Type I Error} = \alpha = P(\text{Rejecting null hypothesis} \mid H_0 \text{ is true}) \quad (6.3)$$

It is important to understand the difference between the  $p$ -value and the significance value  $\alpha$ . Probability value ( $p$ -value) is the evidence for the null hypothesis whereas significance value  $\alpha$  is the error based on repetitive sampling. Hubbard *et al.* (2003) state that the  $p$ -value in a hypothesis test refers to probability of observing the data given a null hypothesis, whereas the significance level  $\alpha$  refers to incorrect rejection of null hypothesis when it is true under **repeated trials**.

2. **Type II Error:** Conditional probability of failing to reject a null hypothesis (or retaining a null hypothesis) when the alternative hypothesis is true is called Type II Error or False Negative (falsely believing that there is no relationship). Usually Type II error is denoted by the symbol  $\beta$ . Mathematically, Type II error can be defined as follows:

$$\text{Type II Error} = \beta = P(\text{Retain null hypothesis} \mid H_0 \text{ is false}) \quad (6.4)$$

The value  $(1 - \beta)$  is known as the power of hypothesis test. That is, the power of the test is given by

$$\text{Power of the test} = 1 - \beta = 1 - P(\text{Retain null hypothesis} \mid H_0 \text{ is false}) \quad (6.5)$$

Alternatively the power of test  $= 1 - \beta = P(\text{Reject null hypothesis} \mid H_0 \text{ is false})$

Description of Type 1 error, Type 2 error, and the power of test is given in Table 6.3.

## 6.5 | HYPOTHESIS TESTING FOR POPULATION MEAN WITH KNOWN VARIANCE: Z-TEST

Z-test (also known as one-sample Z-test) is used when a claim (hypothesis) is made about the population parameter such as population mean or proportion when population variance is known. In this section, we will be discussing the hypothesis testing for the population mean when the population variance is known. Since the hypothesis test is carried out with just one sample, this test is also known as **one-sample Z-test**. According to the central limit theorem (CLT) for sampling distribution of mean, we

**TABLE 6.3** Description of type I error, type II error, and the power of test

Actual Value of $H_0$	Decision made about Null Hypothesis Based on the Hypothesis Test	
	Reject $H_0$	Retain $H_0$
$H_0$ is true	Type I error $P(\text{Reject } H_0 \mid H_0 = \text{true}) = \alpha$	Correct Decision $P(\text{Retain } H_0 \mid H_0 = \text{true}) = (1 - \alpha)$
$H_0$ is false	Correct Decision (Power of test) $P(\text{Reject } H_0 \mid H_0 = \text{false}) = 1 - \beta$	Type II Error $P(\text{Retain } H_0 \mid H_0 = \text{false}) = \beta$

know that the sampling distribution of mean from an independent and identically distributed population for large sample follows a normal distribution with mean  $\mu$  and standard deviation  $\sigma / \sqrt{n}$ . The standardized value  $\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$  follows a standard normal distribution. Z-test uses CLT to conduct a hypothesis test for population mean when the population variance is known; the test statistics for Z-test is given by

$$\text{Z-statistic} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \quad (6.6)$$

The critical value in this case will depend on the significance value  $\alpha$  and whether it is a one-tailed or two-tailed test. The critical value for different values of  $\alpha$  is shown in Table 6.4.

In Excel, the function NORMSINV( $\alpha$ ) [and NORM.S.INV( $\alpha$ )] can be used for finding critical Z-value for left-tailed test. NORMSINV( $1 - \alpha$ ) [and NORM.S.INV( $1 - \alpha$ )] will give the critical Z-value for the right-tailed test. NORMSINV( $\alpha/2$ ) and NORM.S.INV( $1 - \alpha/2$ ) will give critical Z-values for two-tailed test. The decision criteria for rejection or retention of the null hypothesis is described in Table 6.5.

**IMPORTANT**

One-sample Z-test is used when

1. Testing the value of population mean when population standard deviation is known.
2. The population is a normal distribution and the population variance is known.
3. The sample size is large and the population variance is known. That is, the assumption of normal distribution can be relaxed for large samples ( $n > 30$ ).

**TABLE 6.4** Critical value for different values of  $\alpha$ 

$\alpha$	Approximate Critical Values		
	Left-Tailed Test	Right-Tailed Test	Two-Tailed Test
0.1	-1.28	1.28	-1.64 and 1.64
0.05	-1.64	1.64	-1.96 and 1.96
0.01	-2.33	2.33	-2.58 and 2.58

**TABLE 6.5** Condition for rejection of null hypothesis  $H_0$ 

Type of Test	Condition	Decision
Left-tailed test	$Z\text{-statistic} < \text{Critical value}$	Reject $H_0$
	$Z\text{-statistic} \geq \text{Critical value}$	Retain $H_0$
Right-tailed test	$Z\text{-statistic} > \text{Critical value}$	Reject $H_0$
	$Z\text{-statistic} \leq \text{Critical value}$	Retain $H_0$
Two-tailed test	$ Z\text{-statistic}  >  \text{Critical Value} $	Reject $H_0$
	$ Z\text{-statistic}  \leq  \text{Critical Value} $	Retain $H_0$

**EXAMPLE 6.1**

An agency based out of Bangalore claimed that the average monthly disposable income of families living in Bangalore is greater than INR 4200 with a standard deviation of INR 3200. From a random sample of 40,000 families, the average disposable income was estimated as INR 4250. Assume that the population standard deviation is INR 3200. Conduct an appropriate hypothesis test at 95% confidence level ( $\alpha = 0.05$ ) to check the validity of the claim by the agency.

**Solution:**

*In contexts such as this, we set alternative hypothesis as the statement that we would like to prove.*

**Claim:** Average disposable income is more than INR 4200

Let  $\mu$  and  $\sigma$  denote the mean and standard deviation in the population. The corresponding null and alternative hypotheses are

$$\begin{aligned} H_0: \mu &\leq 4200 \\ H_A: \mu &> 4200 \end{aligned}$$

Since we know the population standard deviation, we can use the  $Z$ -test. The corresponding  $Z$ -statistic is given by

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{4250 - 4200}{3200 / \sqrt{40000}} = 3.125$$

This is a right-tailed test. The corresponding  $Z$ -critical value at  $\alpha = 0.05$  for right-tailed test is approximately 1.64 [in Excel  $\text{NORMSINV}(1 - \alpha)$  that is  $\text{NORMSINV}(0.95)$  gives the critical value for the right-tailed test]. Since the calculated  $Z$ -statistic value is greater than the  $Z$ -critical value, we reject the null hypothesis. The corresponding

$p\text{-value} = 0.00088$  [ $p\text{-value}$  in Excel is given by  $1 - \text{NORMSDIST}(Z\text{-statistic value})$ , that is  $1 - \text{NORMSDIST}(3.125)$  in this case]. The critical value,  $Z$ -statistic value, and the corresponding  $p$ -value are shown in Figure 6.5.

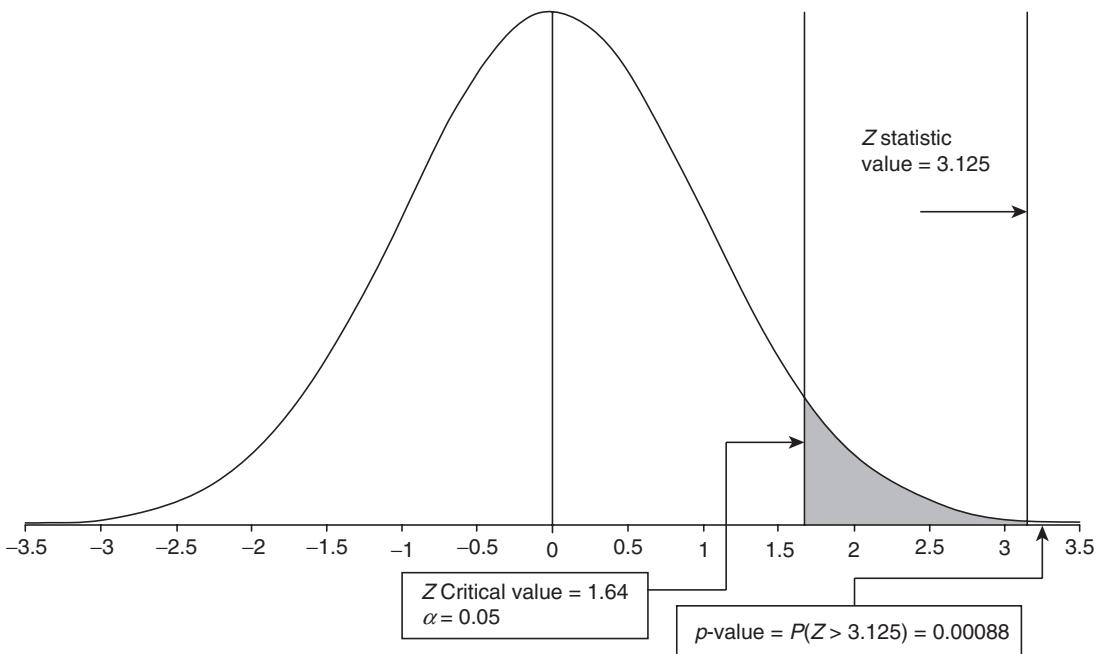


FIGURE 6.5 Critical value,  $Z$ -statistic value, and corresponding  $p$ -value.

**IMPORTANT**

$Z$ -statistic measures the standardized difference between estimated value of mean and the hypothesis value of mean.  $Z = 3.125$  implies that the sample mean is at 3.125 standard deviations away from the hypothesized population mean given that the null hypothesis is true

**EXAMPLE 6.2**

A passport office claims that the passport applications are processed within 30 days of submitting the application form and all necessary documents. Table 6.6 shows processing time of 40 passport applicants. The population standard deviation of the processing time is 12.5 days. Conduct a hypothesis test at significance level  $\alpha = 0.05$  to verify the claim made by the passport office.

**TABLE 6.6** Passport processing time

16	16	30	37	25	22	19	35	27	32
34	28	24	35	24	21	32	29	24	35
28	29	18	31	28	33	32	24	25	22
21	27	41	23	23	16	24	38	26	28

**Solution:**

Null and alternative hypotheses in this case are given by

$$H_0: \mu \geq 30$$

$$H_A: \mu < 30$$

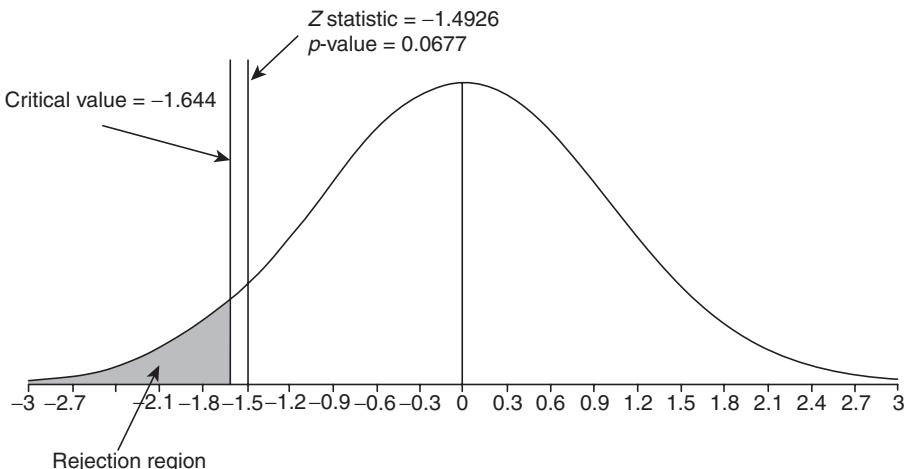
From the data in Table 6.6, the estimated sample mean is 27.05 days.

The standard deviation of the sampling distribution  $\sigma / \sqrt{n} = 12.5 / \sqrt{40} = 1.9764$ .

The value of Z-statistic is given by

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{27.05 - 30}{12.5 / \sqrt{40}} = -1.4926$$

The critical value of left-tailed test for  $\alpha = 0.05$  is  $-1.644$ . Since the critical value is less than the Z-statistic value, we fail to reject the null hypothesis. The  $p$ -value for  $Z = -1.4926$  is 0.06777 which is greater than the value of  $\alpha$ . That is, there is no strong evidence against null hypothesis so we retain the null hypothesis, which is  $\mu \geq 30$ . Figure 6.6 shows the calculated Z-statistic value and the rejection region.

**FIGURE 6.6** Left-tailed test for Example 6.2.

**EXAMPLE 6.3**

According to the company IQ Research, the average Intelligence Quotient (IQ) of Indians is 82 derived based on a research carried out by Professor Richard Lynn, a British Professor of Psychology, using data collected from 2002 to 2006 (Source: IQ Research<sup>1</sup>). The population standard deviation of IQ is estimated as 11.03. Based on a sample of 100 people from India, the sample IQ was estimated as 84.

- (a) Conduct an appropriate hypothesis test at  $\alpha = 0.05$  to validate the claim of IQ Research (that average IQ of Indians is 82).
- (b) Ministry of education believes that the IQ is more than 82. If the actual IQ (population mean) of Indians is 86, calculate the Type II error and the power of hypothesis test.

**Solution:**

(a) Hypothesis test: It is given that  $\mu = 82$ ,  $\sigma = 11.03$ ,  $n = 100$ , and  $\bar{X} = 84$ .

The null and alternative hypotheses in this case are:

$$H_0: \mu = 82$$

$$H_A: \mu \neq 82$$

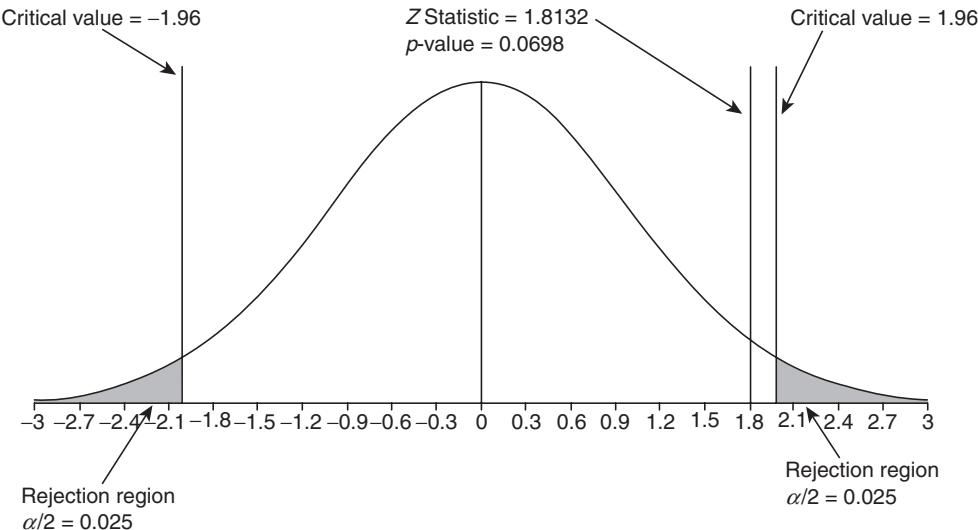
Since the direction of alternative hypothesis is both ways, we have a two-tailed  $t$ -test. The test statistics is given by

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{84 - 82}{11.03 / \sqrt{100}} = 1.8132$$

For a two-tailed test, the critical values at  $\alpha/2 = 0.025$  are  $-1.96$  and  $1.96$  [in Excel  $\text{NORMSINV}(0.025) = -1.96$  and  $\text{NORMSINV}(1 - 0.025) = 1.96$ ]. Since the calculated  $Z$ -statistic value is within the critical values, we fail to reject the null hypothesis (retain the null hypothesis). Figure 6.7 shows the rejection regions and the  $Z$ -statistic value in this case. Since the  $Z$ -statistic value is 1.8132 and falls on the right tail, we first calculate normal distribution beyond 1.8132 which is equal to 0.0348. Since this is a two-tailed test, the  $p$ -value is twice the area to the right side of the  $Z$ -statistic value, which is = 0.0698, that is the  $p$ -value in this case is 0.0698.

---

<sup>1</sup> Source: <https://iq-research.info/en/page/average-iq-by-country>



**FIGURE 6.7** Z-statistic, critical values, and the rejection region for Example 6.3.



In a two-tailed test, the  $p$ -value is two times the tail area.

- (a) **Calculating Type II Error and Power of Test:** In this case, the null and alternative hypotheses are

$$H_0: \mu \leq 82$$

$$H_A: \mu > 82$$

Note that ministry of education believes that the average IQ is 86 (thus we have to carry out a right-tailed test). Type II error is the conditional probability of retaining a null hypothesis when it is false, that is  $P(\text{retaining } H_0 \mid H_0 \text{ is false})$ .

The mean and standard deviation of Z-statistic in null hypothesis are 82 and 1.103, respectively. For the standard normal distribution the critical value for a right tailed test when  $\alpha = 0.05$  is 1.644. The corresponding critical value for the normal distribution  $N(82, 1.103)$  is

$$X_{\text{critical}} = \mu + Z_\alpha \times \sigma / \sqrt{n} = 82 + 1.644 \times 1.103 = 83.8133$$

That is, under normal distribution  $N(82, 1.103)$ , the region beyond 83.8133 is the rejection region (rejection of null hypothesis).

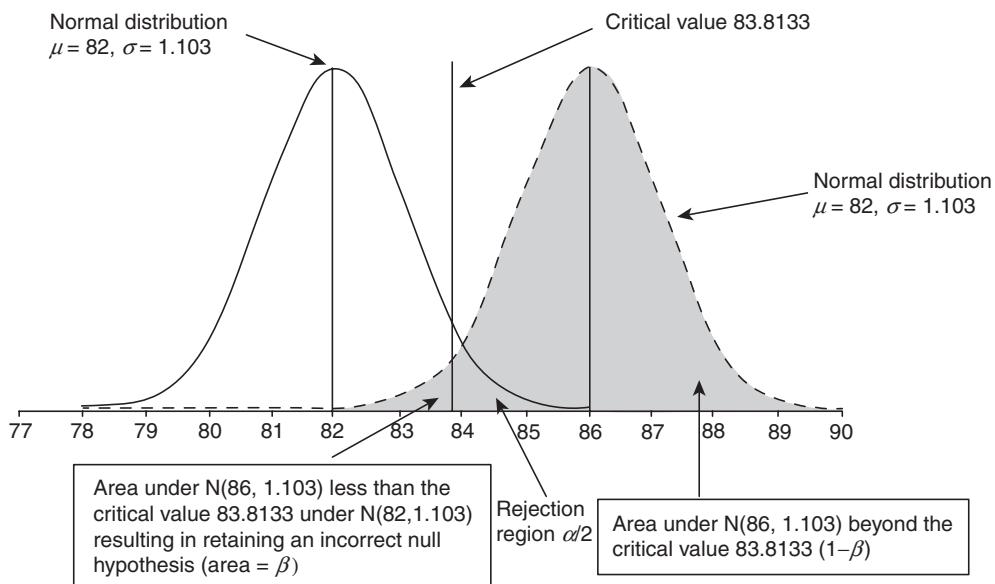
Now consider the normal distribution  $N(86, 1.103)$ . Area under this normal distribution may take values below 83.8133 which is region of retaining the null hypothesis, although the actual mean in this case is 86. Thus, we will be retaining the null hypothesis when it is incorrect resulting in Type II error,  $\beta$  (Figure 6.8).

For the normal distribution  $N(86, 1.103)$ , the probability of the variable taking value less than 83.8133 (the critical value) is given by

$$P(X \leq 83.8133) = P\left(Z \leq \frac{83.8133 - 86}{1.103}\right) = 0.0237$$

That is, the Type II error  $\beta = 0.0237$

The power of test,  $1 - \beta = 1 - 0.0237 = 0.9763$



**FIGURE 6.8** Type II error and power of hypothesis test.

### 6.5.1 | Power of Test and the Power Function

The power of the test  $1 - \beta$  is the conditional probability of rejecting the null hypothesis when the alternative hypothesis is true. For different values of the actual value of population mean, we can calculate the power  $(1 - \beta)$ . The plot between different mean values and  $(1 - \beta)$  is called the **power function** and is shown in Figure 6.9.

Figure 6.9 shows the change in power of test as the actual value of mean changes.

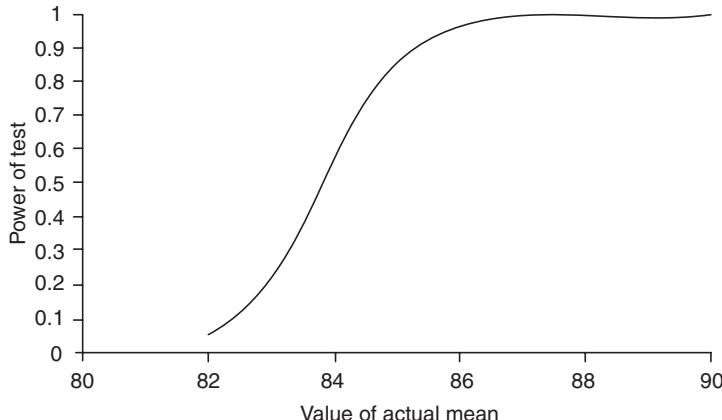


FIGURE 6.9 Power function.

## 6.6 | HYPOTHESIS TESTING FOR POPULATION PROPORTION: Z-TEST FOR PROPORTION

In this section, we will be discussing the hypothesis testing for population proportion based on one sample (and thus called **one-sample test for proportion**). According to the central limit theorem of proportions, the sampling distribution of proportions  $\hat{p}$  for a large sample follows an approximate normal distribution with mean  $p$  (the population proportion) and standard deviation  $\sqrt{\frac{p(1-p)}{n}}$ . That is, the Z-statistic as defined below will follow a standard normal distribution:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \quad (6.7)$$

To calculate the standard deviation  $\sqrt{\frac{p(1-p)}{n}}$  we need the knowledge of  $p$ . However, we can use the value of  $\hat{p}$  estimated from large samples. One of the thumb rules used is that the value of  $n \times \hat{p} \times (1 - \hat{p}) \geq 10$  to use Eq. (6.7). Few authors also suggest that  $n \times \hat{p} \times (1 - \hat{p})$  should be at least 15 so that we can calculate the population standard deviation from the estimated value of proportion from the sample. Since, the statistic in this case is Z-statistic, the critical values are same as in the case of one-sample test for population mean with known variance.

### EXAMPLE 6.4

According to a study exactly 12% of gift cards purchased from e-commerce portals are never used. The manager of an e-commerce company wanted to test whether this claim is true. She collected data of 250 gift card purchases and found that 22 gift cards were not used till its expiry date.

- (a) Conduct an appropriate hypothesis test at 5% significance to check whether the claim that exactly 12% gift cards are never used is true or not.
- (b) Calculate the 95% confidence interval for the proportion of gift cards that are not used.

**Solution:**

- (a) The estimated value of proportion of gift cards not used is

$$\hat{p} = \frac{22}{250} = 0.088$$

$$n \times \hat{p} \times (1 - \hat{p}) = 250 \times 0.088 \times (1 - 0.088) = 20.064 > 10$$

so we can use  $\hat{p}$  to calculate the population standard deviation.

The initial claim is that the percentage of unused gift cards is equal to 12%. The null and alternative hypotheses are

$$H_0: p = 0.12$$

$$H_A: p \neq 0.12$$

The value of the Z-statistic is given by

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.088 - 0.12}{\sqrt{\frac{0.12 \times (1 - 0.12)}{250}}} = -1.557$$

Note that the critical values are  $-1.96$  and  $1.96$  (two-tailed test). Since the calculated value of  $Z$  is not part of the rejection region (greater than  $-1.96$  and less than  $1.96$ ), we retain the null hypothesis that  $p = 0.12$ , the corresponding  $p$ -value is  $0.1195$ . In Excel,  $\text{NORM.S.DIST}(-1.557) = 0.05973$ . This is a two-tailed test, and thus the  $p$ -value is  $2 \times 0.05973 \approx 0.1195$ .

- (b) The 95% confidence interval for the proportion is given by

$$\left[ \hat{p} \pm Z_{\alpha/2} \times \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}} \right] = \left[ 0.088 \pm 1.96 \times \sqrt{\frac{0.088 \times (1 - 0.088)}{20}} \right] = [0.0528, 0.1231]$$

From the confidence interval estimate for the proportion of unused gift cards we can infer that the proportion of unused gift is likely to lie between 0.0528 and 0.1231 at 95% confidence level.

## 6.7 | HYPOTHESIS TEST FOR POPULATION MEAN UNDER UNKNOWN POPULATION VARIANCE: t-TEST

We use the fact that a sampling distribution of a sample from a population that follows normal distribution with unknown variance follows a  $t$ -distribution with  $(n - 1)$  degrees of freedom. In many cases the population variance (and thus the standard deviation) will not be known. In such cases we will have to estimate the variance using the sample itself. Let  $S$  be the standard deviation estimated from the sample of size  $n$ . Then the statistic  $\frac{\bar{X} - \mu}{S/\sqrt{n}}$  will follow a  $t$ -distribution with  $(n - 1)$  degrees of freedom if the

sample is drawn from a population that follows a normal distribution. Here 1 degree of freedom is lost since the standard deviation is estimated from the sample. Thus, we use the  $t$ -statistic (hence the test is called  $t$ -test) to test the hypothesis when the population standard deviation is unknown.

$$t\text{-statistic} = \frac{\bar{X} - \mu}{S/\sqrt{n}} \quad (6.8)$$

### IMPORTANT

*The  $t$ -test is used when the population follows a normal distribution and the population standard deviation  $\sigma$  is unknown and is estimated from the sample.  $t$ -test is a robust test for violation of normality of the data as long as the data is close to symmetry and there are no outliers.*

### EXAMPLE 6.5

Aravind Productions (AP) is a newly formed movie production house based out of Mumbai, India. AP was interested in understanding the production cost required for producing a Bollywood movie. The industry believes that the production house will require at least INR 500 million (50 crore) on average. It is assumed that the Bollywood movie production cost follows a normal distribution. Production cost of 40 Bollywood movies in millions of rupees are shown in Table 6.7. Conduct an appropriate hypothesis test at  $\alpha = 0.05$  to check whether the belief about average production cost is correct.

**TABLE 6.7** Production cost of Bollywood movies

601	627	330	364	562	353	583	254	528	470
125	60	101	110	60	252	281	227	484	402
408	601	593	729	402	530	708	599	439	762
292	636	444	286	636	667	252	335	457	632

### Solution:

It is given that the production cost of Bollywood movies follows a normal distribution; however, the standard deviation of the population is not known and we need

to estimate the standard deviation value from the sample. Thus, we have to use the *t*-test for testing the hypothesis. From the sample data in Table 6.7 we get the following values:

$$n = 40, \bar{X} = 429.55, \text{ and } S = 195.0337$$

The null and alternative hypotheses are

$$H_0: \mu \leq 500$$

$$H_A: \mu > 500$$

The corresponding test statistic is

$$\text{t - statistic} = \frac{\bar{X} - \mu}{S / \sqrt{n}} = \frac{429.55 - 500}{195.0337 / \sqrt{40}} = -2.2845$$

Note that this is a one-tailed test (right-tailed) and the critical *t*-value at  $\alpha = 0.05$  under right-tailed test,  $t_{\text{critical}} = 1.6848$  [in Excel  $\text{TINV}(2\alpha, df)$  will return right-tailed critical value at significance of  $\alpha$ , in this example  $\alpha = 0.05$ , the corresponding critical *t*-value using Excel function is  $\text{TINV}(0.1, 39) = 1.6848$ , that is the critical value is 1.6848]. Since *t*-statistic value is less than the critical *t*-value, we retain the null hypothesis. The *t*-statistic value and critical value for the *t*-test are shown in Figure 6.10.

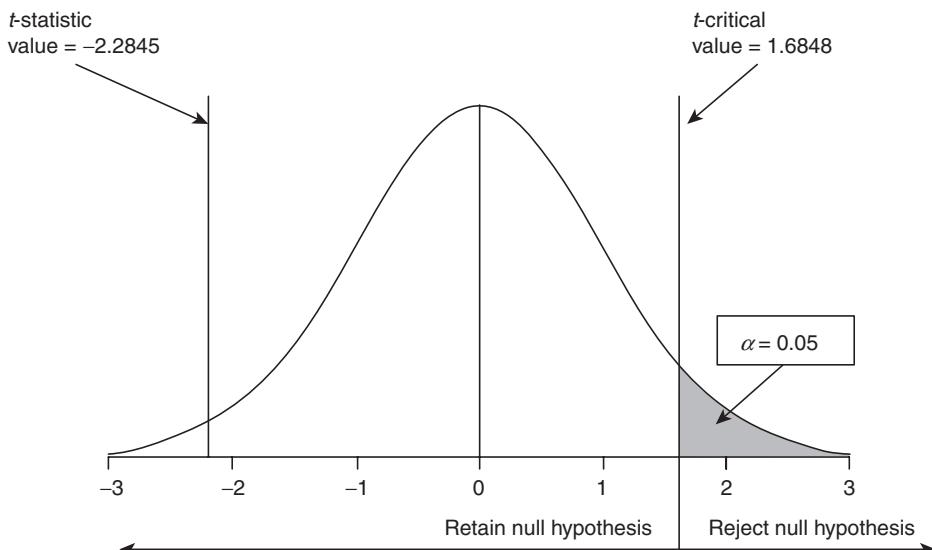


FIGURE 6.10 Critical value, *t*-statistic value for *t*-test in Example 6.5.

## EXAMPLE 6.6

According to statistics released by the Department of Civil Aviation, the average delay of flights is equal to 16.8 minutes, flight delays are assumed to follow a normal distribution. However, from a sample of 50 flights, the average delay was estimated to be 19.5 minutes and the sample standard deviation was 6.6 minutes. Conduct a hypothesis test to disprove the claim that the average delay is equal to 16.8 minutes at  $\alpha = 0.01$ .

**Solution:**

Given  $n = 50$ ,  $\bar{X} = 19.5$ ,  $S = 6.6$

Null and alternative hypotheses are

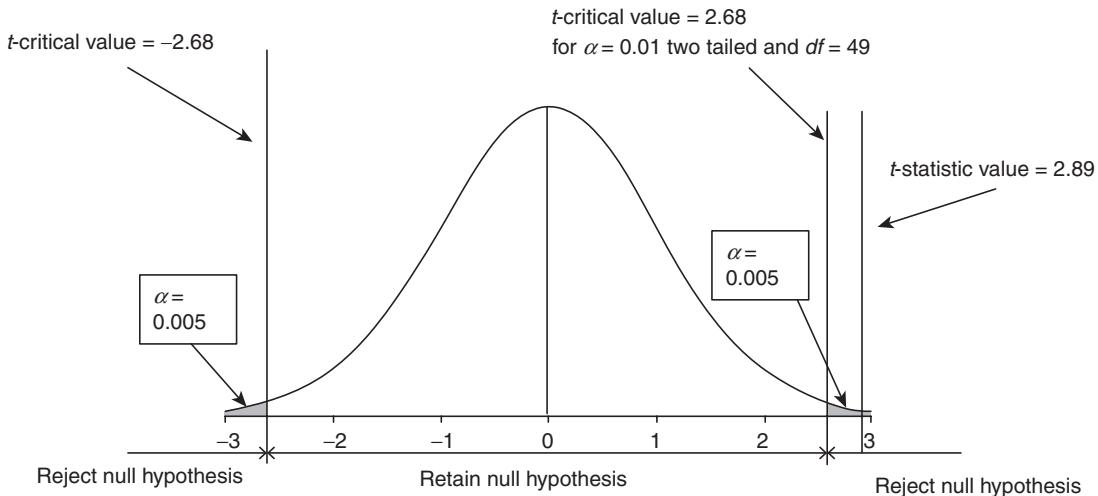
$$H_0: \mu = 16.8$$

$$H_A: \mu \neq 16.8$$

The corresponding  $t$ -statistic value is

$$t = \frac{\bar{X} - \mu}{S / \sqrt{n}} = \frac{19.5 - 16.8}{6.6 / \sqrt{50}} = 2.8927$$

The critical  $t$ -value for two-tailed  $t$ -test when  $\alpha = 0.01$  and degrees of freedom = 49 is 2.67 [in Excel,  $\text{TINV}(0.01, 49) = 2.68$  or  $\text{T.INV.2T}(0.01, 49) = 2.68$ ]. Since the calculated  $t$ -statistic value is greater than the  $t$ -critical value, we reject the null hypothesis. The corresponding  $p$ -value is 0.0057 [in excel  $\text{T.DIST}(t\text{-statistic value, degrees of freedom, tails})$  returns the  $p$ -value,  $\text{T.DIST.2T}(2.8927, 49) = 0.0057$ ]. The values of  $t$ -statistic,  $t$ -critical value, rejection and retention regions are shown in Figure 6.11.



**FIGURE 6.11**  $t$ -statistic,  $t$ -critical, rejection and acceptance regions for Example 6.6.

## 6.8 | PAIRED SAMPLE t-TEST

In many cases, we would like to analyse whether an intervention (or treatment) such as training programs, marketing promotions, treatment for specific illness, and life style changes may have significantly changed the population parameter values such as mean and proportion before and after the intervention. The objective in this case is to check whether the difference in the parameter values is statistically significant before and after the intervention or between two different types of interventions (for example two different types of promotions). In a paired *t*-test, the data related to the parameter is captured twice from the same subject, once before the intervention and once after intervention. Alternatively, the paired *t*-test can be used for comparing two different interventions such as two different promotion strategies applied on the same subject (price discount versus bundling of items). Examples of paired *t*-test are as follows:

1. Body weight of subjects before and after attending a yoga training program.
2. Cholesterol levels of subjects before and after attending meditation training.
3. Amount of time spent by subjects on the internet before and after marriage.
4. Quantity of alcohol consumed by people before and after breakup.
5. Level of cortisol among students during and after exam.

Note that, in the above examples, we are observing a population parameter value on the same subject before and after intervention. Assume that the mean difference in the estimated parameter value before and after the treatment is  $D$ , and the corresponding standard deviation of difference is  $S_d$ . Let  $\mu_d$  be the hypothesized mean difference. Then the statistic defined in Eq. (6.9) follows a *t*-distribution with  $(n - 1)$  degrees of freedom.

$$\frac{D - \mu_d}{S_d / \sqrt{n}} \quad (6.9)$$

Here we assume that the differences follow a normal distribution.

### EXAMPLE 6.7

Table 6.8 shows data on alcohol consumption before and after breakup. Conduct a paired *t*-test to check whether the alcohol consumption is more after the breakup (that is  $\mu_d > 0$ ) at 95% confidence ( $\alpha = 0.05$ ).

**TABLE 6.8** Average weekly consumption of alcohol (in ml) before and after breakup

S. No.	Before Breakup ( $X_1$ )	After Breakup ( $X_2$ )	Difference ( $X_2 - X_1$ )
1	470	408	-62
2	354	439	85
3	496	321	-175
4	351	437	86

**TABLE 6.8** Average weekly consumption of alcohol (in ml) before and after breakup—Continued

S. No.	Before Breakup ( $X_1$ )	After Breakup ( $X_2$ )	Difference ( $X_2 - X_1$ )
5	349	335	-14
6	449	344	-105
7	378	318	-60
8	359	492	133
9	469	531	62
10	329	417	88
11	389	358	-31
12	497	391	-106
13	493	398	-95
14	268	394	126
15	445	508	63
16	287	399	112
17	338	345	7
18	271	341	70
19	412	326	-86
20	335	467	132

The mean difference, that is mean of  $(X_2 - X_1)$ , is 11.5 and the corresponding sample standard deviation is 95.67.

The null and alternative hypotheses are (when the claim is that the difference is greater than zero):

$$H_0: \mu_d \leq 0$$

$$H_A: \mu_d > 0$$

The value of test statistic is

$$t = \frac{D - \mu_d}{S_d / \sqrt{n}} = \frac{11.5 - 0}{95.6757 / \sqrt{20}} = 0.5375$$

The critical  $t$ -value for one-tailed test when  $\alpha = 0.05$  and  $df = 19$  is 1.7291 [in Excel T.INV(0.05, 19) = 1.7291]. Since the  $t$ -statistic value is 0.5375, which is less than the critical value, we retain the null hypothesis and conclude the difference in alcohol consumption is not greater than 0 before and after breakup. The corresponding  $p$ -value is 0.70.

**EXAMPLE 6.8**

A researcher believes that people drink more coffee on Mondays than other days of the week. Based on a sample of 50 coffee drinkers, the mean difference was estimated as 14 ml and the corresponding standard deviation was 8.5 ml. Conduct an appropriate hypothesis test at  $\alpha = 0.1$  to check the claim that people drink on average 10 ml more coffee on Mondays compared to other days of the week.

**Solution:**

We are given  $n = 50$ ,  $D = 14$ ,  $S_d = 8.5$ , and  $\mu_d = 10$ . The null and alternative hypotheses are

$$H_0: \mu_d \leq 10$$

$$H_A: \mu_d > 10$$

The test statistic is

$$t = \frac{D - \mu_d}{S_d / \sqrt{n}} = \frac{14 - 10}{8.5 / \sqrt{50}} = 3.3275$$

Critical value of  $t$  for  $\alpha = 0.1$  and  $df = 49$  is 1.2990 [in Excel T.INV(1 - 0.1, 49) = 1.2990]. Since  $t$ -statistic value is greater than  $t$ -critical value, we reject the null hypothesis. That is, there is evidence from the data that people drink at least 10 ml more coffee on Mondays than other days. The corresponding  $p$ -value is 0.000834 [in Excel, T.DIST.RT(3.3275, 49) = 0.000834].

## 6.9 | COMPARING TWO POPULATIONS: TWO-SAMPLE Z- AND t-TEST

In many cases we would like to compare parameters of two different populations to check for any difference in the parameter values such as mean. In this section, we will be discussing 3 scenarios and the corresponding test statistic for comparing two population means using two-sample Z-tests and  $t$ -tests.

### 6.9.1 | Difference in Two Population Means when Population Standard Deviations are Known: Two-Sample Z-Test

In this case we make the following assumptions:

1. The sample sizes (say  $n_1$  and  $n_2$ ) of two samples drawn from two populations are large (say at least 30) and the corresponding standard deviations  $\sigma_1$  and  $\sigma_2$  are known.
2. The samples are drawn from two normally distributed populations with corresponding standard deviations  $\sigma_1$  and  $\sigma_2$  known.

Assume that  $\mu_1$  and  $\mu_2$  are the population means. Our interest is to check a hypothesis on difference between  $\mu_1$  and  $\mu_2$ , that is  $(\mu_1 - \mu_2)$ . If  $\bar{X}_1$  and  $\bar{X}_2$  are the estimated mean values from two samples drawn

from two populations, the statistic  $(\bar{X}_1 - \bar{X}_2)$  follows a standard normal distribution with mean  $(\mu_1 - \mu_2)$  and standard deviation  $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ , where  $n_1$  and  $n_2$  are the sample sizes of two samples. The corresponding Z-statistic is given by

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (6.10)$$



*Z-statistic in Eq. (6.10) is valid only when the population standard deviations are known and the samples are from a normal distribution.*

#### EXAMPLE 6.9

The Dean of St Peter School of Management Education (SPSME) believes that the graduating students with specialization in Marketing earn at least INR 5000 more per month than the students with specialization in Operations Management. To verify his belief, the Dean collected a sample data from his graduating students, given in Table 6.9. Conduct an appropriate hypothesis test at  $\alpha = 0.05$  to check whether the difference in monthly salary is at least 5000 more for students with marketing specialization compared to operations specialization. Assume that the salary of students with marketing specialization and operations specialization follow normal distribution.

**TABLE 6.9** Sample values on marketing and operations students

Specialization	Sample Size	Estimated Mean Salary (in Rupees) per Month	Population Standard Deviation
Marketing	120	67,500.00	7,200
Operations	45	58,950.00	4,600

**Solution:**

We have  $n_1 = 120$ ,  $n_2 = 45$ ,  $\bar{X}_1 = 67,500$ ,  $\bar{X}_2 = 58,950$ ,  $\sigma_1 = 7,200$  and  $\sigma_2 = 4,600$ . The null and alternative hypotheses are

$$H_0: \mu_1 - \mu_2 \leq 5000$$

$$H_A: \mu_1 - \mu_2 > 5000$$

The corresponding test statistic value is

$$Z = \frac{(67500 - 58950) - 5000}{\sqrt{\frac{7200^2}{120} + \frac{4600^2}{45}}} = \frac{3550}{949.85} = 3.7374$$

The critical value of  $Z$  at  $\alpha = 0.05$  is 1.64 [= NORMSINV(1 – 0.05)]. Since the  $Z$ -statistic value is higher than the  $Z$ -critical value, we reject the null hypothesis. The corresponding  $p$ -value is  $9.29 \times 10^{-5}$ .

### 6.9.2 | Difference in Two Population Means when Population Standard Deviations are Unknown and Believed to be Equal: Two-Sample t-Test

In this section we discuss the hypothesis test for difference in two population means when the standard deviations of the populations are unknown. Hence we need to estimate them from the samples drawn from these two populations. An additional assumption we make here is that the standard deviation of two populations are equal (however, unknown). Then the sampling distribution of the difference in estimated means ( $\bar{X}_1 - \bar{X}_2$ ) follows a  $t$ -distribution with  $(n_1 + n_2 - 2)$  degrees of freedom with mean  $(\mu_1 - \mu_2)$  and standard deviation

$$\sqrt{S_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} \quad (6.11)$$

where  $S_p^2$  is the pooled variance of two samples and is given by

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)} \quad (6.12)$$

The corresponding  $t$ -statistic is

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad (6.13)$$

#### EXAMPLE 6.10

A company makes a claim that children (in the age group between 7 and 12) who drink their health drink will grow taller than the children who do not drink that health drink. Data in Table 6.10 shows average increase in height over one-year period from two groups: one drinking the health drink and the other not drinking the health drink. At  $\alpha=0.05$ , test whether the increase in height for the children who drink the health drink is at least 1.2 cm.

**TABLE 6.10** Data related in increase in heights from different groups

Group	Sample Size	Increase in Height (in cm) during the Test Period	Standard Deviation Estimated from Sample
Drink health drink	80	7.6 cm	1.1 cm
Do not drink health drink	80	6.3 cm	1.3 cm

**Solution:**

We have  $n_1 = 80$ ,  $n_2 = 80$ ,  $\bar{X}_1 = 7.6$ ,  $\bar{X}_2 = 6.3$ ,  $\sigma_1 = 1.1$ , and  $\sigma_2 = 1.3$ . (Note that the sample standard deviations need not be same for samples although the population standard deviations are same.)

The null and alternative hypotheses are

$$H_0: \mu_1 - \mu_2 \leq 1.2$$

$$H_A: \mu_1 - \mu_2 > 1.2$$

Pooled variance is

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)} = \frac{79 \times 1.1^2 + 79 \times 1.3^2}{80 + 80 - 2} = 1.45$$

The  $t$ -statistic is

$$t = \frac{(7.6 - 6.3) - 1.2}{\sqrt{1.45 \left( \frac{1}{80} + \frac{1}{80} \right)}} = 0.5252$$

The  $t$ -critical value for one-tailed  $t$ -test when  $\alpha = 0.05$  and degrees of freedom = 158 ( $80 + 80 - 2$ ) is 1.6546. Since the calculated  $t$ -statistic value is less than  $t$ -critical value we retain the null hypothesis. That is, the difference between two groups is less than 1.2 and the corresponding right-tailed test has a  $p$ -value of 0.3.

### 6.9.3 | Difference in Two Population Means when Population Standard Deviations are Unknown and Not Equal: Two-Sample t-Test with Unequal Variance

In this section we discuss the hypothesis test for difference in two population means, when the standard deviations of the two populations are unknown and unequal. We need to estimate standard deviations from the samples drawn from these two populations. Then the sampling distribution of the difference in estimated means ( $\bar{X}_1 - \bar{X}_2$ ) follows a  $t$ -distribution with mean ( $\mu_1 - \mu_2$ ) and standard deviation

$$S_u = \sqrt{\left( \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)} \quad (6.14)$$

The corresponding degrees of freedom is given by

$$df = \left\lfloor \frac{\frac{S_u^4}{(\bar{S}_1^2/n_1)^2 + (\bar{S}_2^2/n_2)^2}}{\frac{n_1-1}{n_1} + \frac{n_2-1}{n_2}} \right\rfloor \quad (6.15)$$

where the symbol  $\lfloor \rfloor$  implies rounding down to the nearest integer. The  $t$ -statistic for testing two populations with unequal variance is given by

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\left( \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)}} \quad (6.16)$$

### EXAMPLE 6.11

A researcher is interested in finding the average duration of marriage based on the educational qualifications of couples. Two groups were considered for the study: Group 1 consisted of couples with no Bachelor's degree (both partners) and Group 2 consisted of couple who both have Bachelor's degree or higher. Data in Table 6.11 shows average duration of marriage in years. At  $\alpha = 0.05$ , test whether the average duration of marriage is more for couples with no Bachelor's degree as compared to couples with Bachelor's degree.

**TABLE 6.11** Data related to duration of marriage from different groups

Group	Sample Size	Duration of Marriage in Years	Standard Deviation Estimated from Sample
Couples with no degree	120	10.1 years	2.4 years
Couples with degree	100	9.5 years	3.1 years

**Solution:**

We have  $n_1 = 120$ ,  $n_2 = 100$ ,  $\bar{X}_1 = 10.1$ ,  $\bar{X}_2 = 9.5$ ,  $\sigma_1 = 2.4$ , and  $\sigma_2 = 3.1$ . The null and alternative hypotheses are

$$H_0: \mu_1 - \mu_2 \leq 0$$

$$H_A: \mu_1 - \mu_2 > 0$$

The  $t$ -statistic value is

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}} = \frac{(10.1 - 9.5) - 0}{\sqrt{\frac{2.4^2}{120} + \frac{3.1^2}{100}}} = 1.5805$$

The corresponding degrees of freedom is

$$df = \left\lfloor \frac{S_u^4}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}} \right\rfloor = \left\lfloor \frac{0.0207}{0.000113} \right\rfloor = \lfloor 184.33 \rfloor = 184$$

The critical value of  $t$  for  $\alpha = 0.05$  and  $df = 184$  is 1.6531. Since the  $t$ -statistic is less than critical value of  $t$ , we retain the null hypothesis. That is, the difference in duration of marriage between two groups is less than or equal to zero. The corresponding  $p$ -value is 0.05785.

## 6.10 | HYPOTHESIS TEST FOR DIFFERENCE IN POPULATION PROPORTION UNDER LARGE SAMPLES: TWO-SAMPLE Z-TEST FOR PROPORTIONS

When the proportions are estimated from a large sample, then sampling distribution of proportions follows a normal distribution according to the central limit theorem. Let  $\hat{p}_1$  and  $\hat{p}_2$  be the estimated values of proportions from large samples.

The difference  $\hat{p}_1 - \hat{p}_2$  is the hypothesized difference in population proportions. When the null hypothesis is  $H_0: p_1 = p_2$  (that is,  $H_0: \hat{p}_1 = \hat{p}_2$ ), then the test statistic is given by

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1 - \hat{p}) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad (6.17)$$

where  $\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$  is the pooled estimate for proportion.

### EXAMPLE 6.12

The marketing manager of a company believes that the non-affluent customers are sensitive to discounts compared to affluent customers. To validate this hypothesis, discount coupons were sent to non-affluent and affluent customers and the data is provided in Table 6.12. Use an appropriate hypothesis to check whether there is any difference in proportion of customers who use discount coupons at  $\alpha = 0.05$ .

**TABLE 6.12** Data related to increase in heights from different groups

Group	Sample Size	Number of Customers using Discount Coupons	Estimated Proportion
Non-Affluent	500	145	0.29
Affluent	300	42	0.14

**Solution:**

The null and alternative hypothesis are

$$H_0: p_1 = p_2$$

$$H_A: p_1 \neq p_2$$

From Table 6.12, the estimated values of proportions are  $\hat{P}_1 = 0.29$  and  $\hat{P}_2 = 0.14$ . The pooled proportion is (assuming null hypothesis is true)

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = \frac{500 \times 0.29 + 300 \times 0.14}{500 + 300} = 0.2338$$

The test statistic is

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1 - \hat{p}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{(0.29 - 0.14) - 0}{\sqrt{0.2338 \left( \frac{1}{500} + \frac{1}{300} \right)}} = \frac{0.15}{0.03091} = 4.8528$$

The  $Z$ -critical at  $\alpha = 0.05$  for two-tailed test is 1.96. Since the calculated value of  $Z$  is more than the  $Z$ -critical value, we reject the null hypothesis. That is, the non-affluent customers are sensitive to coupons, that is they use more coupons (which can be verified using one tailed test).

## 6.11 | EFFECT SIZE: COHEN'S D

Effect size is a measure of magnitude or strength of relationship between two or more groups in a population; the greater the effect size, the greater will be the power of test. Cohen (1977) defined effect size for various statistical tests. For testing the difference between two sample means, Cohen suggested the following equation to calculate the effect size:

$$\text{Cohen's D} = \frac{|\bar{X}_1 - \bar{X}_2|}{\sigma} \quad (6.18)$$

where  $\sigma$  is the pooled standard deviation. Cohen's D values of 0.2, 0.5, and 0.8 are considered as small, medium, and large effect sizes. These values correspond to proportions of explained variance of 1%,

5.9%, and 13.8% (Sawyer and Ball, 1981). Cohen's D is useful in variable or feature selection. A variable with high Cohen's D implies existence of relationship between two variables.

## 6.12 | HYPOTHESIS TEST FOR EQUALITY OF POPULATION VARIANCES

In this section, we will discuss hypothesis test for checking whether two population variances are equal. Recall from Chapter 5 that if  $S^2$  is the variance estimated from a sample of size  $n$  drawn from a population that follows a normal distribution, then the ratio  $\frac{(n-1)S^2}{\sigma^2}$  follows a  $\chi^2$  distribution with  $(n - 1)$  degrees of freedom. That is,

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \quad (6.19)$$

Rearranging Eq. (6.19), we get

$$S^2 = \frac{\chi^2 \sigma^2}{n-1} \quad (6.20)$$

Let  $S_1^2$  and  $S_2^2$  be the variances estimated from two samples drawn from two different populations that follow normal distribution. Then the ratio of variances [using Eq. (6.20)] is given by

$$\frac{S_1^2}{S_2^2} = \frac{\chi_1^2 \sigma_1^2 / (n_1 - 1)}{\chi_2^2 \sigma_2^2 / (n_2 - 1)} \quad (6.21)$$

If the variances are equal ( $\sigma_1^2 = \sigma_2^2$ ), Eq. (6.21) is given by

$$\frac{S_1^2}{S_2^2} = \frac{\chi_1^2 / (n_1 - 1)}{\chi_2^2 / (n_2 - 1)} \quad (6.22)$$

Equation (6.22) is a ratio of two chi-square distributions divided by the corresponding degrees of freedom, which is an  $F$ -distribution with  $(n_1 - 1)$  degrees of freedom for numerator and  $(n_2 - 1)$  degrees of freedom for the denominator. The test statistic for testing the equality of variances is given by

$$F_{(n_1-1, n_2-1)} = \frac{S_1^2}{S_2^2} \quad (6.23)$$

Note that since  $F$ -distribution is not a symmetrical distribution, we have to calculate both left and right critical values separately in case of a two-tailed test. However, it is usually carried as a right-tailed test.

### EXAMPLE 6.13

Preetha Dallal is an investment advisor and she believes that the variance of stock prices of manufacturing companies and information technology companies are the same. To verify the claim, variances of stock prices from these two sectors are collected and are shown in Table 6.13. Conduct an appropriate test at  $\alpha = 0.1$  to check whether the variances of stock prices in two industry sectors are equal or not.

**TABLE 6.13** Data related in variance in stock prices from different groups

Group	Sample Size	Variance Estimated from Sample
Manufacturing Firms	80	42
Information Technology Firms	52	36

**Solution:**

Given  $S_1^2 = 42$  and  $S_2^2 = 36$ . The null and alternative hypotheses are

$$H_0: \sigma_1^2 = \sigma_2^2$$

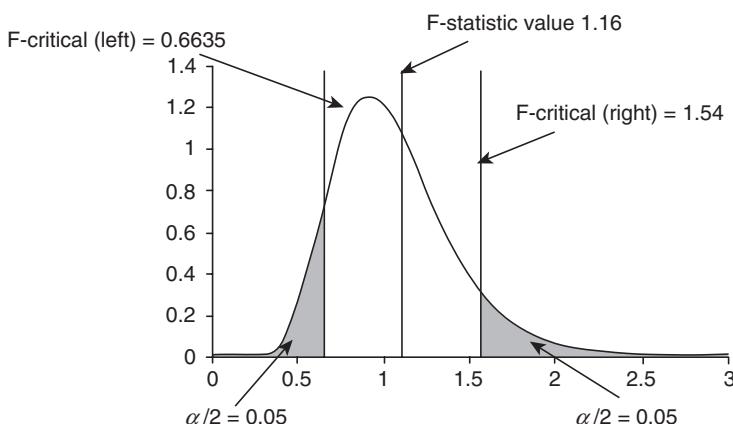
$$H_A: \sigma_1^2 \neq \sigma_2^2$$

The corresponding  $F$ -statistic is given by

$$F_{(n_1-1, n_2-1)} = F_{(79, 51)} = \frac{42}{36} = 1.1666$$

The left and right critical values are given by 0.6635 and 1.5407, respectively. Since the calculated value is between the two critical values, we will retain the null hypothesis. In Excel, the right critical value of the  $F$ -test is given by  $\text{FINV. RT}(\alpha/2, df-N, df-D)$ , where  $df-N$  is the degrees of freedom in numerator and  $df-D$  is the degrees of freedom in the denominator. The left critical value in Excel is given by  $(1/\text{FINV}(\alpha/2, df-D, df-N))$ . That is, we reverse the degrees of freedom and calculate  $\text{FINV}$  and take the inverse to get the left critical value under  $F$ -test.

Figure 6.12 shows the  $F$ -distribution with left and right critical values and the calculated  $F$ -statistic value.

**FIGURE 6.12**  $F_{(79, 51)}$  distribution with left and right critical values.

An alternative and frequently used approach to test the variances of two groups is through right tailed  $F$  test to simplify the calculations. In such cases, the null and alternative hypotheses are set as

$$H_0: \sigma_1^2 \leq \sigma_2^2$$

$$H_A: \sigma_1^2 > \sigma_2^2$$

While performing a right-tailed test, we choose sample with larger variance as  $S_1^2$ . In Example 6.13, we will designate  $S_1^2 = 42$  and  $S_2^2 = 36$ . So the  $F$ -statistic for right-tailed test is

$$F_{(79, 51)} = \frac{42}{36} = 1.1666$$

The critical  $F$ -value for the right-tailed test is 1.39 [in Excel,  $\text{FINV}(\alpha, df_1, df_2)$  or  $\text{F.INV.RT}(\alpha, df_1, df_2)$  functions can be used for getting the  $F$ -critical value,  $df_1$  and  $df_2$  are degrees for freedom for numerator and denominator, respectively]. Since the calculated  $F$ -statistic value is less than the  $F$ -critical value, we retain the null hypothesis.

## 6.13 | NON-PARAMETRIC TESTS: CHI-SQUARE TESTS

In the previous sections, we discussed methods of testing hypothesis which are about population parameters and make certain assumptions about the population distribution. We assumed that the test statistic follows standard normal distribution,  $t$ -distribution, or  $F$ -distribution. Tests such as  $Z$ -test,  $t$ -test, and  $F$ -test are called parametric tests since the objective is to infer about a population parameter such as mean and proportion in case of single sample tests or compare population parameters in the case of two sample tests. To conduct  $Z$ -test and  $t$ -test we need summary statistics (mean and/or standard deviation), and not necessarily the entire distribution. In this section, we will be discussing non-parametric tests (also known as distribution free tests since they do not have assumptions about distribution of the population). *Non-parametric tests* imply that the tests are not based on the assumptions that the data is drawn from a probability distribution defined through parameters such as mean, proportion and standard deviation.

A major difference between parametric and non-parametric tests is that in a parametric test we need only values of the parameter and the knowledge about the distribution, whereas in case of non-parametric test we use the entire distribution of the data. Importantly, the data may not follow any parametric distribution such as normal distribution. Also, the test is not about the population parameter but about characteristics of the entire distribution (for example, whether the data follows a normal distribution or not). A non-parametric method for hypothesis tests is used when one or more of the following conditions exist in the test:

1. The test is not about the population parameter such as mean and standard deviation.
2. The method does not require assumptions about population distribution (such as population follows normal distribution).

### 6.13.1 | Chi-Square Goodness of Fit Tests

Goodness of fit tests are hypothesis tests that are used for comparing the observed distribution of data with expected distribution of the data to decide whether there is any statistically significant difference between the observed distribution and a theoretical distribution (such as exponential, normal, Weibull, etc.) based on comparison of observed frequencies in the data and the expected frequencies if the data follows a specified theoretical distribution.

The null and alternative hypotheses in chi-square goodness of fit tests are

$H_0$ : There is no statistically significant difference between the observed frequencies and the expected frequencies from a hypothesized distribution.

$H_A$ : There is a statistically significant difference between the observed frequencies and the expected frequencies from a hypothesized distribution.

Let  $Z$  be a standard normal distribution (that is  $Z = \frac{X_i - \mu}{\sigma}$ ). Then the random variable  $Z^2$  follows a chi-square distribution with 1 degree of freedom (since  $X_i$  is the only random variable and there are no constraints). If we have  $k$  random variables, namely,  $X_1, X_2, \dots, X_k$ , then a chi-square distribution with  $k$ -degrees of freedom is given by

$$\chi^2(k) = \left( \frac{X_1 - \mu}{\sigma} \right)^2 + \left( \frac{X_2 - \mu}{\sigma} \right)^2 + \dots + \left( \frac{X_k - \mu}{\sigma} \right)^2 \quad (6.24)$$

If we replace the population mean  $\mu$  with sample mean, then the degrees of freedom will be  $(k - 1)$ . Even if the population is not normally distributed, the right-hand side of the equation will follow an approximate  $\chi^2$  distribution for a large sample.

Consider a binomial random variable with parameter  $p$  (probability of success) and number of trials  $n$ . Then for a large sample, the standardized random variable in Eq. (6.25) follows a standard normal distribution (central limit theorem for proportions):

$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \quad (6.25)$$

Note that  $\hat{p} = \frac{X_1}{n}$ . Substituting this in Eq. (6.25), we get

$$Z = \frac{\frac{X_1}{n} - p}{\sqrt{p(1-p)/n}} = \frac{X_1 - np}{\sqrt{np(1-p)}} \quad (6.26)$$

$Z^2$  is given by

$$Z^2 = \frac{(X_1 - np)^2}{np(1-p)} \quad (6.27)$$

Let  $X_2 = n - X_1$ . Then Eq. (6.27) is equivalent to

$$Z^2 = \frac{(X_1 - np)^2}{np(1-p)} = \frac{(X_1 - np)^2}{np} + \frac{[X_2 - n(1-p)]^2}{n(1-p)} \quad (6.28)$$

Proof of Eq. (6.28) may be found in advanced statistical books on chi-square test (Elston and Johnson, 2008). Note that  $np$  and  $n(1 - p)$  are the expected values of two categories (success and failure) of the binomial distribution. The two parts on the right-hand side of Eq. (6.28) are of form

$$\frac{(\text{Observed value} - \text{Expected value})^2}{\text{Expected value}} \quad (6.29)$$

Equation (6.29) can be generalized to distributions with multiple groups. Thus, the chi-square statistic for goodness of fit test is given by

$$\chi^2 \text{ statistic} = \sum_{i=1}^n \sum_{j=1}^m \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (6.30)$$

where  $O_{ij}$  is the observed frequency in category  $(i, j)$  and  $E_{ij}$  is the expected frequency in the category  $(i, j)$ . The numerator in Eq. (6.30) is zero when observed and expected frequencies are same (starting value of Chi-square distribution); when the difference is large, the numerator will be a large number falling on the right side of the distribution. Thus, chi-square test is always a right-tailed test. In goodness of fit tests, degrees of freedom is  $k - c - 1$ , where  $k$  is the number of groups,  $c$  is the number of parameters estimated from the data. Ideally, expected frequencies in each group should be at least 5.

#### EXAMPLE 6.14

Hanuman Airlines (HA) operated daily flights to several Indian cities. One of the problems HA faces is the food preferences by the passengers. Captain Cook, the operations manager of HA, believes that 35% of their passengers prefer vegetarian food, 40% prefer non-vegetarian food, 20% low calorie food, and 5% request for diabetic food. A sample of 500 passengers was chosen to analyse the food preferences and the data is shown in Table 6.14. Conduct a chi-square test to check whether Captain Cook's belief is true at  $\alpha = 0.05$ .

**TABLE 6.14** Sample preferences of 500 customers

Food Type	Vegetarian	Non-Vegetarian	Low Calorie	Diabetic
Number of Passengers	190	185	90	35

#### Solution:

The null and alternative hypotheses in this case are given as

$H_0$ : Probability distribution of the food preference is  $P(\text{Vegetarian}) = 0.35$ ;  
 $P(\text{Non-Vegetarian}) = 0.40$ ;  $P(\text{Low Calorie}) = 0.20$ , and  $P(\text{Diabetic}) = 0.05$

$H_A$ : Probability distribution of the food preference is not as defined in null hypothesis

Since the sample is 500, we can calculate the expected values for various food preferences using the proportions given in the question. Table 6.15 shows the observed, expected frequencies, and the chi-square statistic calculations.

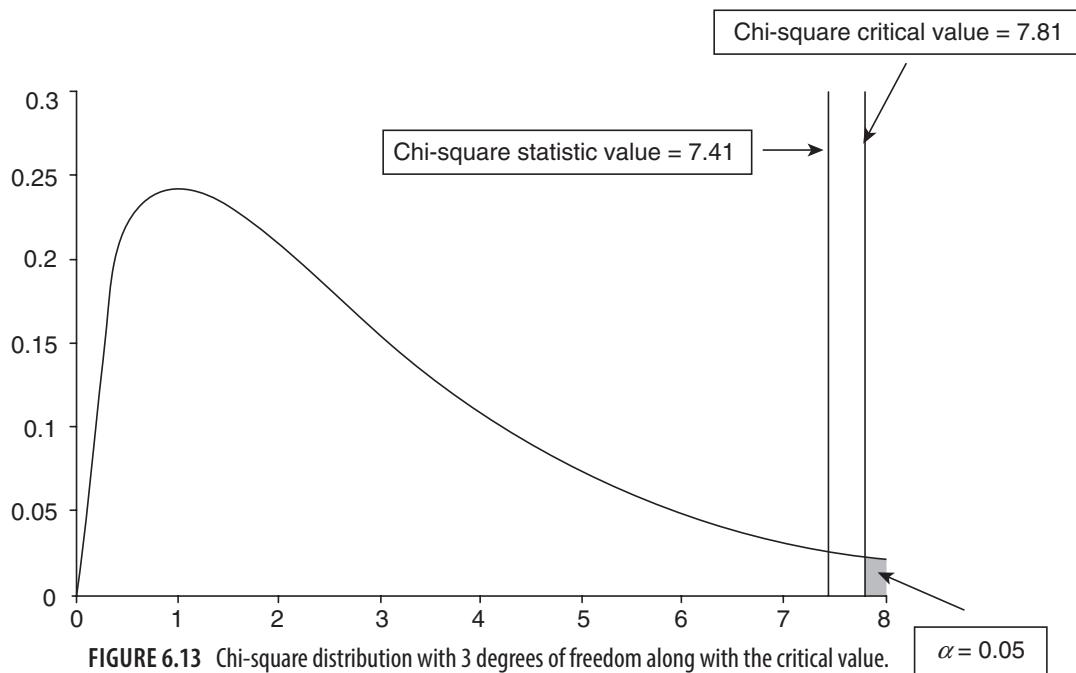
**TABLE 6.15** Observed frequencies, expected frequencies, and chi-square statistics

Food Type	Observed Frequency ( $O_i$ )	Expected Frequency ( $E_i$ )	$\frac{(O_i - E_i)^2}{E_i}$
Vegetarian	190	175	1.285
Non-Vegetarian	185	200	1.125
Low Calorie	90	100	1
Diabetic	35	25	4

The chi-square statistic value is given by

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 1.285 + 1.125 + 1 + 4 = 7.410$$

Note that this is a right-tailed test. The chi-square critical value ( $df = 4 - 1 = 3$ ) for  $\alpha = 0.05$  is 7.8147. Since the calculated chi-square value is less than the critical value we retain the null hypothesis. That is, we conclude that Captain Cook's belief about the food preferences of HA is true. Figure 6.13 shows the critical values and the chi-square statistic values.



**EXAMPLE 6.15**

Kachara Seth is a general manager of the ‘Clean City & Co’ that manages solid wastes produced in different cities across India. He believes that the quantity of solid waste produced in Bangalore is 5000 tons per day with a standard deviation of 300 tons. Kachara Seth collected data on solid waste produced on randomly selected days. The data (on 50 days) is shown in Table 6.16. Check whether the data actually follows a normal distribution at  $\alpha = 0.01$ .

**TABLE 6.16** Solid waste data

4640	4967	4640	4967	4640
4957	5169	4957	5169	4957
5064	5033	5064	5033	5064
5062	4514	5062	4514	5062
5217	4883	5217	4883	5217
4658	4998	4658	4998	4658
5557	4843	5557	4843	5557
5510	5112	5510	5112	5510
5005	5111	5005	5111	5005
4967	4865	4967	4865	4967

**Solution:**

The null and alternative hypotheses in this case are

$H_0$ : There is no difference between the observed frequencies and expected frequencies from a normal distribution with mean 5000 and standard deviation 300.

$H_A$ : There is a difference between the observed frequencies and the expected frequencies under normal distribution with mean 5000 and standard deviation 300.

To check whether the data follows a normal distribution or not, we have to create groups and check the observed and expected frequencies within the group. The number of groups,  $N$ , can be calculated using the following equation when the sample size is  $n$  (Sturges, 1926):

$$N = \lfloor 1 + 3.3 \log_{10}(n) \rfloor \quad (6.31)$$

The data on solid waste is collected over 50 days, thus  $n = 50$ . The number of groups is

$$N = \lfloor 1 + 3.3 \log_{10}(n) \rfloor = \lfloor 1 + 3.3 \log_{10}(50) \rfloor = \lfloor 6.6 \rfloor = 6$$

The width of the group can be now calculated using the following formula:

$$\text{Range} = \frac{\text{Max} - \text{Min}}{N} = \frac{5557 - 4514}{6} = 173.83$$

The range of the groups may be rounded to 174 to create the 6 groups as shown in Table 6.17. The observed frequencies in each group ( $O_i$ ) can be easily obtained from Table 6.16. The expected frequencies are calculated using the cumulative distribution function of the normal distribution. For example, let  $L_i$  and  $U_i$  be the lower and upper limits of the group  $i$ . Then the expected frequency for group  $i$  is given by

$$E_i = n \times [F(U_i) - F(L_i)] \quad (6.32)$$

where  $F(L_i)$  and  $F(U_i)$  are the cumulative distribution functions with mean 5000 and standard deviation 300.

**TABLE 6.17** Groups for the data in Table 6.16

Group	Range of the Group	Observed Frequency ( $O_i$ )	Expected Frequency ( $E_i$ )	$\frac{(O_i - E_i)^2}{E_i}$
1	4514	4688	8	4.83
2	4689	4863	2	8.70
3	4864	5038	19	11.26
4	5039	5213	12	10.47
5	5214	5388	3	6.99
6	5389	5563	6	3.35

The chi-square statistic value is given by

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 15.86$$

The chi-square critical value for  $\alpha = 0.05$  (here the degrees of freedom is 5) is 11.07. Since the chi-square statistic value is greater than the chi-square critical value, we reject the null hypothesis and conclude that the data does not follow a normal distribution with mean 5000 and standard deviation 300.

#### EXAMPLE 6.16

Peter Phonewala is the chief operating officer at Airmobile, a mobile phone service provider based out of Coimbatore, India. Peter was interested in finding the probability distribution of the call duration. His friend Erlang Phonewala suggested Peter Phonewala that call duration is most likely to be an exponential distribution.

To check whether the call duration actually follows an exponential distribution, Peter collected a sample of 50 calls and the duration of call in minutes is shown in Table 6.18. Using goodness of fit test, check whether the data follows an exponential distribution.

**TABLE 6.18** Sample call duration (in minutes) data

2.47	4.23	5.41	3.49	4.17	10.09	18.78	0.68	2.28	16.16
0.28	2.97	4.01	5.88	20.32	26.88	19.07	0.22	6.37	10.38
4.2	10.17	1.84	21.88	9.42	0.01	6.15	4.99	3.07	18.6
1.54	10.23	3.99	6.17	0.39	11.03	9.38	1.57	6.91	2.49
5.52	11.53	7.64	8.8	7.17	3.26	6.74	16.32	10	7.45

### Solution:

The null and alternative hypotheses are

$H_0$ : There is no difference between the observed frequencies and expected frequencies from an exponential distribution.

$H_A$ : There is a difference between the observed frequencies and expected frequencies from an exponential distribution.

We can group the data into 6 groups; the minimum and maximum values of the call duration are 0.01 and 26.88. The range is 26.87 and the length of the interval is 4.48. Table 6.19 provides the observed and expected frequencies and chi-square statistic calculation. For exponential distribution, the expected frequency is given by

$$E_i = n \times [\exp(-\lambda \times L_i) - \exp(-\lambda \times U_i)]$$

where the scale parameter  $\lambda = 1/\text{mean call duration}$ . In this case the mean call duration is 7.652 and the corresponding value of  $\lambda = 0.1306$ . The observed, expected, and chi-square statistic calculations are shown in Table 6.19.

**TABLE 6.19** Observed, expected, and chi-square statistic values

Group	Group Range	$O_i$	$E_i$	$\frac{(O_i - E_i)^2}{E_i}$
1	0.01	4.49	20	22.08
2	4.5	8.98	13	12.30
3	8.99	13.47	9	6.85
4	13.48	17.96	2	3.82
5	17.97	22.45	5	2.13
6	22.46	26.94	1	1.18

The chi-square statistic value is given by

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 5.7$$

The chi-square critical value for  $\alpha = 0.05$  (here the degrees of freedom is 4) is 9.48. Since the chi-square statistic value is less than the chi-square critical values, we retain the null hypothesis and conclude that the data follows an exponential distribution.

### 6.13.2 | Choice of Number of Intervals in Chi-Square Goodness of Fit Test

One of the decisions one has to take while conducting a chi-square goodness of fit test is the choice of the number of intervals. Equation (6.30) for number of bins is originally created for finding optimal bins in a histogram. Mann and Wald (1942) suggested the following equation to calculate the optimal number of bins for conducting a chi-square goodness of fit test:

$$k = 4 \left( \sqrt{\left( \frac{2(N-1)^2}{c^2} \right)} \right)^{1/5} \quad (6.33)$$

where  $k$  is the number of bins (intervals),  $N$  is the sample size, and  $c$  is the critical value under standard normal distribution for a given significance value  $\alpha$ .

### 6.13.3 | Chi-Square Test of Independence

Chi-square test of independence is a hypothesis test in which we test whether two (or more) groups are statistically independent or not. For example, assume that a telecom company is interested in checking whether or not the customer churn depends on the customer segment. Here the customers are classified either as churned or retained. Consider the data shown in Table 6.20, which provides customer segment wise the churned and retained customers. Table 6.20 is often called as the **contingency table**.

In Table 6.20, we have observed values for both churned and retained customers. We can calculate the expected frequencies for two classes (churned and retained) for different customer segments using

**TABLE 6.20** Customer churn data (contingency table)

Customer Segment	Churned	Retained	Total
Segment 1	25	250	275
Segment 2	41	484	525
Segment 3	28	172	200
Total	94	906	1000

the basic concept of independent events. From basic theory of probability, we know that if two events  $A$  and  $B$  are independent then  $P(A \cap B) = P(A) P(B)$ .

Let  $O_{ij}$  = Observed number of cases in customer segment  $i$  ( $i = 1, 2, 3$ ) and classification  $j$  (1 = churned and 2 = retained)

That is,  $O_{11}$  is observed number of churned customers in customer segment 1 (which is 25 as per the table). We can calculate the expected number of cases for each of the observed values using the following logic:

Let  $E_{ij}$  = Expected number of cases in customer segment  $i$  ( $i = 1, 2, 3$ ) and classification  $j$  (1 = churned and 2 = retained)

To calculate  $E_{ij}$ , we have to first calculate  $P(i, j)$  [=  $P(i \cap j)$ ], that is the joint probability of  $(i, j)$ . Let  $i$  = segment 1 and  $j$  = churned. Then assuming  $i$  and  $j$  to be independent we can write

$$P(\text{segment 1, churned}) = P(\text{segment 1}) \times P(\text{churned}) = (275/1000) \times (94/1000)$$

That is, for segment = 1 and classification = 1 (churned),

$$\begin{aligned} E_{11} &= n \times P(\text{segment 1}) \times P(\text{churned}) \\ &= 1000 \times (275/1000) \times (94/1000) = (275 \times 94) / 1000 \end{aligned}$$

Note that,

$$E_{11} = (\text{Sum of row 1} \times \text{Sum of column 1}) / \text{Total sum}$$

In general, the value of

$$E_{ij} = (i^{\text{th}} \text{ Row sum} \times j^{\text{th}} \text{ Column sum}) / \text{Total sum}$$

The chi-square test statistic for test of independence is same as the chi-square statistic for goodness of fit test. Generic null and alternative hypotheses for chi-square test of independence are given by

$H_0$ : The variables are independent

$H_A$ : The variables are dependent

For the data in Table 6.20, the null and alternative hypotheses are

$H_0$ : Customer segments and customer churn are independent

$H_A$ : Customer segments and customer churn are dependent

We can calculate the expected values for each segment and churn classification combination. The statistic for chi-square test of independence is given by

$$\chi^2 = \sum_{i=1}^n \sum_{j=1}^m \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

The corresponding degrees of freedom is  $(r - 1) \times (c - 1)$ , where  $r$  is the number of rows and  $c$  is the number of columns in the contingency table. The chi-square statistic calculation for the data in Table 6.20 is shown in Table 6.21.

**TABLE 6.21** Chi-square statistic calculation

Customer Segment	Class	$O_{ij}$	$E_{ij} = \frac{\text{Row sum} \times \text{Column sum}}{\text{Total sum}}$	$= \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
1 ( $i = 1$ )	Churned ( $j = 1$ )	25	25.85	0.02795
2 ( $i = 2$ )	Churned ( $j = 1$ )	41	49.35	1.412817
3 ( $i = 3$ )	Churned ( $j = 1$ )	28	18.8	4.502128
1 ( $i = 1$ )	Retained ( $j = 2$ )	250	249.15	0.0029
2 ( $i = 2$ )	Retained ( $j = 2$ )	484	475.65	0.146584
3 ( $i = 3$ )	Retained ( $j = 2$ )	172	181.2	0.467108

The chi-square statistic value is given by

$$\chi^2 = \sum_{i=1}^3 \sum_{j=1}^2 \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 6.55$$

The chi-statistic value is 6.55 and the critical value is 5.99 (at  $\alpha = 0.05$ ). Thus we reject the null hypothesis, and conclude that the customer segment and customer churn are not independent.

### SUMMARY

1. Hypothesis testing is the basis for inferential statistics since in many cases we may not have access to the entire population and inference about the population parameter has to be made based on a sample.
2. Hypothesis testing is an integral part of predictive analytics algorithms such as multiple linear regression (MLR) and logistic regression (LR). The model selection (variable selection or feature selection) in MLR and LR is achieved through hypothesis testing.
3. Hypothesis is a claim and hypothesis testing is a process used to either reject or retain the claim.
4. Hypothesis testing involves four important steps: 1. Identification of null and alternative hypothesis, 2. Identification test statistic, 3. Calculation of  $p$ -value, and 4. Decision criteria for rejection or retention of null hypothesis.
5. The central limit theorem is used to derive the test statistic in the case of  $Z$  and  $t$  tests.
6. The probability value or  $p$ -value is the evidence in support of the null hypothesis.  $P$ -value gives the conditional probability of observing the test statistic value given the null hypothesis is true.
7. Decision to reject or retain a null hypothesis is taken by comparing the  $p$ -value with the significance value  $\alpha$ . Significance value  $\alpha$  is the conditional probability of rejecting a null hypothesis when the null hypothesis is true. The value of  $\alpha$  is the Type I error.
8. Retaining a null hypothesis when it is false is called Type II error, usually denoted by  $\beta$ . The value of  $1 - \beta$  is called the power of the hypothesis test.
9. Hypothesis testing can be broadly classified into parametric testing and non-parametric testing. In a parametric hypothesis testing, the test is about a population parameter, whereas in the case of non-parametric testing the testing is about not the population parameter. For example, in the case of chi-square goodness of fit test, the test is for the entire distribution of data.
10. Hypothesis testing is probably one of the most important concepts in analytics since it forms the basis for many predictive analytics algorithms.

**MULTIPLE CHOICE QUESTIONS**

1. The  $p$ -value in a hypothesis test is
  - Probability of rejecting the null hypothesis.
  - Probability of accepting the null hypothesis.
  - Probability of observing the test statistic value when the null hypothesis is true.
  - Probability of observing the test statistic value when the null hypothesis is false.
2. Which of the following statements about null hypothesis are correct?
  - Null hypothesis is complement of alternative hypothesis.
  - Null hypothesis will consist of equality sign.
  - At the beginning the null hypothesis is assumed to be true.
  - Null hypothesis will not include the equality sign.
3. The significance value  $\alpha$  is
  - Probability accepting a true null hypothesis.
  - Conditional probability retaining a null hypothesis when null hypothesis is true.
  - Probability accepting an alternative hypothesis.
  - Conditional probability of rejecting a null hypothesis when null hypothesis is true.
4. The one-sample Z-test is used when
  - The sample is drawn from a normal distribution.
  - When the sample size is large.
  - When the sample size is large and the standard deviation is estimated from the sample.
  - When the sample size is large and the population standard deviation is known.
5. The critical values of Z-test for  $\alpha = 0.05$  under left-, right- and two-tailed tests are
  - $-1.64, +1.64, |1.96|$
  - $-1.96, +1.96, |1.64|$
  - $+1.64, -1.64, |1.96|$
  - $+1.96, -1.96, |1.64|$
6. In a hypothesis test
  - The objective is to reject null hypothesis
  - The objective is to retain alternative hypothesis
  - The objective is to decide whether to reject or retain null hypothesis
  - The objective is to decide whether to reject or retain alternative hypothesis
7. The power of hypothesis test is (when  $\alpha$  = Type I error and  $\beta$  = Type II error)
  - $1 - \alpha$
  - $\alpha$
  - $\beta$
  - $1 - \beta$
8. A researcher is interested in conducting a hypothesis test on the mean value of population. If the standard deviation is estimated from the sample, the appropriate hypothesis test is
  - One-sample  $t$ -test
  - One-sample Z-test
  - Two-sample  $t$ -test
  - Chi-square test
9. In a right-tailed Z-test, the  $p$ -value is
  - $P(Z \geq Z\text{-statistic})$
  - $P(Z \leq Z\text{-statistic})$
  - $P(Z \geq \alpha)$
  - $P(Z \leq \alpha)$
10. An HR manager is interested in checking the effectiveness of a training program on the sales team. The appropriate test for checking the effectiveness of the training program is:
  - Two population  $t$ -test with equal variance
  - Paired  $t$ -test
  - Two population  $t$ -test with unequal variance
  - Chi-square test of independence

**EXERCISES**

- David Daruvala is a Manager at Madhuveloka, a chain of pubs located in Bangalore. David believes that average weekly consumption of beer among his customers is at least 1200 ml. From a sample of 45 customers he estimated the sample average as 1220 ml and the standard deviation estimated from the sample is 120 ml. Use an appropriate hypothesis test at  $\alpha=0.01$  to check whether the average consumption of beer is at least 1200 ml. Calculate the  $p$ -value and confidence interval for the average consumption of beer.
- Peter Poulouse is the Vice President of an e-commerce company called ‘We Sell Everything On Earth (WSEOE)’. Peter believes that at least 12% of WSEOE’s customers return the products purchased by them. To validate his belief, he took a sample of 620 customers and found that 80 customers had returned the products. Carry out an appropriate hypothesis test at a significance of 0.1% to check whether the return is at least 12%.
- Ammajon.com is an e-commerce company with headquarters in Bangalore and sells millions of merchandize across India. One of the problems they encounter is fraud; it is believed that e-commerce fraud in India is at least 6%. Ammajon’s analytics team collected a random sample of 1200 cases and found 66 fraudulent transactions. Conduct a hypothesis test and check whether the fraud proportion at Ammajon is at least 6% at  $\alpha=0.01$  and calculate the corresponding  $p$ -value and confidence interval for the proportion of fraud.
- At Die Another Day (DAD) hospital the average daily food wastage is 22 kg with a standard deviation of 4 kg based on data collected over 30 days. Dr Who, the operations manager at DAD hospital, implements few lean and six sigma projects to reduce the food wastage. After the implementation, the wastage was 21 kg and standard deviation was 2.5 kg based on a sample of 30 days. Do an appropriate hypothesis test at 90% confidence level ( $\alpha=0.10$ ) to check whether the lean and six sigma projects have reduced the food wastage.
- Hedge funds are alternative investment options that claim to provide better returns compared to investments such as mutual funds and stocks. Hedge funds use many strategies such as *Convertible*, *Currency*, *Derivative*, *Emerging Market*, etc. Siddharth Sinha, an investment advisor at Platinum Investments, strongly believes that the average returns of the hedge fund strategy ‘*Emerging Market*’ is higher than that of ‘*Derivative*’. A sample of hedge funds that uses these two strategies and their returns are shown in Tables 6.22 and 6.23. Conduct a hypothesis test to check whether the strategy ‘*Emerging Market*’ gives more average returns than the strategy ‘*Derivatives*’.

**TABLE 6.22** Percentage returns of hedge funds under strategy ‘*Emerging Market*’

11.20	12.10	13.33	16.40	15.00	10.00	12.00	13.00	12.00	13.00
8.25	7.00	10.00	11.46	11.00	7.70	7.00	12.00	18.00	10.00
13.11	9.00	14.00	9.90	16.00	9.00	6.00	11.40	7.00	16.00
8.41	17.21	14.00	15.00	17.20	18.00	9.00	7.00	15.45	15.00
13.00	18.60	16.00	9.60	12.00	6.00	15.00	8.00	16.29	9.00

**TABLE 6.23** Percentage returns of hedge funds under strategy ‘*Derivatives*’

17.65	10.20	19.00	14.00	11.00	4.97	11.00	7.00	5.12	4.90
19.00	11.45	16.00	6.87	14.00	8.00	10.78	16.00	18.00	11.00
13.00	17.00	18.00	16.00	12.00	13.26	19.00	10.00	17.00	5.56
8.00	15.55	11.22	6.78	10.00	19.00	14.00	15.00	14.00	7.00
14.00	15.00	18.00	7.78	10.00	15.00	16.20	15.00	11.65	13.00

- Average stock price of automotive sector is believed to be less than the average stock price of banking and finance sector. To validate this belief, share prices of 40 automotive companies and share prices of

30 companies from banking and finance sector were collected and are given in Tables 6.24 and 6.25. Conduct a hypothesis test at  $\alpha = 0.05$  to check whether the average share price of automobile companies is lesser than that of banking and finance companies.

**TABLE 6.24** Share values of 40 companies from automotive sector

261.88	489.47	301.02	408.39	245.42	593.42	293.41	388.39	348.80	380.51
442.18	492.64	288.30	490.88	272.60	564.17	613.13	319.60	587.93	532.34
482.77	236.20	262.41	243.24	633.97	267.15	633.43	603.53	342.55	426.23
646.58	653.45	215.17	408.42	651.18	301.00	282.25	471.30	468.73	481.38

**TABLE 6.25** Share values of 30 companies from banking and finance sector

469.78	343.39	226.30	237.55	515.78	613.14	441.19	644.43	386.82	597.71
288.33	491.84	291.82	632.43	431.54	571.88	321.87	598.11	319.14	233.19
617.78	292.50	178.80	452.48	428.82	242.80	461.33	424.19	342.12	168.34

7. At Kumbakonam Cooperative Bank (KCB) loan defaults (non-performing assets) account for 10%. Krishnan Iyer, Manager at KCB, believes that the proportion of defaults is dependent on the profession of the applicants since certain professions have higher probability of job loss. A sample data was collected from the past loan and a contingency table was created as shown in Table 6.26. Check whether the profession and loan defaults are dependent at  $\alpha = 0.01$ .

**TABLE 6.26** Loan default contingency table

Profession	Loan Status		
	Defaults	Non-Default	Total
Lawyers	60	160	220
Government Service	80	320	400
Management at Private Companies	45	220	265
Total	185	700	885

8. Beautiful Looks (BL) is a chain of health and beauty care hospitals that specializes in plastic surgery. A sample of surgery costs is collected to understand the distribution of the surgery cost and is shown in Table 6.27. Check at  $\alpha = 0.05$ , whether the surgery costs follow a normal distribution.

**TABLE 6.27** Sample surgery costs from 60 past surgeries

15076	17516	11119	14598	12725	16043	16667	17004	18096	13345
17422	12365	16523	16374	15527	12341	11614	16319	19937	14664
16404	16869	15711	13065	16385	11750	15550	17734	11830	11557
18008	17700	15047	12938	13986	10411	11881	9991	13476	14514
14309	13978	18529	14174	15693	16992	13266	15962	14347	11901
13397	13687	14314	14429	13678	14738	13280	17007	13407	19102

9. Data about time between failures of an avionic system was collected and is shown in Table 6.28. The time between failures for avionic systems is believed to follow an exponential distribution. Conduct a hypothesis test at  $\alpha = 0.01$  to check whether the data follows an exponential distribution.

**TABLE 6.28** Time to failure data

577.28	28.31	1260.8	406.80	273.57	90.46	123.41	8.52	14.85	17.97
133.32	637.49	214.99	254.66	17.23	368.27	7.07	220.54	558.76	203.42
149.93	670.16	943.00	312.91	141.03	1164.1	125.13	90.19	302.09	108.85
303.94	53.05	19.49	65.74	31.53	73.05	496.25	523.91	21.78	605.28
130.74	319.80	340.84	725.46	52.48	49.03	1241.7	495.51	438.29	639.06

## Case Study

### Central Parking Services Private Limited<sup>2</sup>

Howards End mall is an important site for us and we need to predict traffic trends of vehicles to the mall accurately so that we can plan our pricing and resources accordingly. Otherwise, we will never be able to have an effectively managed workforce and the pricing resulting in customer dissatisfaction.

— Poornima, Director for New Initiatives, Central Parking Services

Although these thoughts were reverberating in her mind for quite some time, it was during the board meeting in March 2013 that Poornima decided to voice her thoughts. Central Parking Services (CPS) with headquarters located in Bangalore, India provided parking solutions at various malls, office buildings, airports, residential apartments, hospitals, and so on throughout India (shown in **Exhibit 1**). The discussion was about frequent operational problems they faced such as demand supply gaps for parking bays, longer waiting times at the entry and exits. The recent series of such events at one of their biggest sites 'Howards End' mall was an example, where owing to sudden incoming traffic of cars, there were frequent alarms, resulting in many complaints from unsatisfied customers. Howards End<sup>3</sup> was a mall situated in one of the most posh areas in the heart of Mumbai, the financial capital of India. It harboured all the top retail brands available in India and was laced with other essentials such as a 400-seater food court and a 500-seater cinema hall. This site witnessed a huge spike in traffic especially during weekends where the workforce faced several challenges in managing the parking space.

Most of the workforce at Howards End comprised permanently hired staff and was managed on a shift system. One of the solutions, which were proposed, was keeping extra staff during weekends. However, the incoming traffic was unpredictable and employing an optimal manpower still remained a challenge. According to Poornima:

<sup>2</sup> The case study is authored with Tanmay Gupta and Abhishek Srivastava and Professor U Dinesh Kumar and is distributed through Harvard Business Publishing. Copyright © the Indian Institute of Management Bangalore and reproduced with permission of IIM Bangalore. The case is not intended to serve as an endorsement, source of primary data, or to show effective handling of decision or process.

<sup>3</sup> Name of the mall changed for confidentiality purpose.

**Continued...**

We deploy more staff at this mall during weekends, to maintain smooth flow of vehicles in and out of the parking lot. The operational cost increases during weekends, so we need to use dynamic pricing considering the demand and supply of parking lot as well as increase in the staff.

CPS invested approximately INR 12,00,000 (1\$ = INR 62, in November 2013) per parking bay in any retail mall to make it functional. So, it was extremely important that its manpower requirement was met and the parking fee pricing was right. CPS followed a thumb rule for planning manpower which was also influenced by factors such as type of technology used and design of the building. If the parking lot had facilities such as an automated ticketing facility, then the manpower needed was lesser. The number of employees assigned to a parking lot also depended upon estimates of seasonality of demand. CPS deployed more manpower during weekends, national holidays, and other religious festivals. The strategy also changed based on the region that CPS operated in. For example, in Kolkata, the festival of Dussehra/Durga Pooja generally celebrated in October attracted large crowds, whereas in cities down south such as Chennai, the festival of Pongal celebrated in January was considered more important. CPS also looked at the structure of the parking lot in order to plan manpower.

### Central parking services – the origins

The origins of CPS can be traced to Building Control Solutions started in 1996 as a building management system company. Sathya and Poornima started Building Control Solutions with personal capital. Sathya was the Chief Executive Officer (CEO), Amit was the Chief Operating Officer (COO), and Venkatesan was the Chief Finance Officer (CFO). Poornima served as the Director for new innovations in the company. The vision with which the company started was to spread its presence across the country since the company was the first mover in the business of building management system in India. During 1996–2001, it was doing projects for building management systems along with home automation. During the initial days, the company was into installing and maintaining temperature and humidity control systems. After about 5 years, the company decided to diversify into providing parking solutions. Its clients wanted CPS to work with the architects to design the parking lot, be involved with equipment planning as well as manage staff requirements of the parking lot. While the company was started in Bangalore, it secured projects across the country. First, the projects were only limited to metropolitan cities, but increasingly the company also secured more and more projects in tier 2 and tier 3 cities. CPS also expanded its presence to shopping malls, hospitals, airports, metros, etc. Besides expanding geographically, the company has been completely focused on its value proposition of being a technological leader in this industry. Since the technology in this industry is constantly changing, CPS has also forged technological partnerships with companies from Germany, Japan, and China.

### Strategy for success in the organized parking industry

Expansion had occurred quickly for CPS. The auto industry in India was booming and so was the retail industry. From this stemmed a demand for organized parking spaces across the ever-expanding urban

**Continued...**

landscape of the country owing to increase in the number of new malls and apartments. There were a few cities in India such as Bangalore, Chennai, Delhi, and Mumbai where the surge in demand was immense. This was fueled by growth in various manufacturing and service companies operating within these cities. It is in these cities where CPS started its operations, and with the meteoric rise of these cities, the company's scale of business also grew. CPS was present at the right place at the right time.

It had taken full benefit of the first-mover advantage and was at a strong position in the market; however, the fast growth also created a few problems. Although, the standards of technology that the company maintained were equivalent, the service practices as well as operational practices differed from time to time. While Poornima headed new initiatives, she was also very closely involved in customer relationship management as well as the operations. Therefore, these discrepancies in operations were not hidden from her.

Poornima said:

The market for automated parking is estimated at INR 150 crore (1 crore = 10 million) in 2013 and this market is projected to grow at 50–60% in the next 5 years. The market is likely to reach INR 500 crore in the next 3–4 years.

The famous Forum mall situated at Koramangala, Bangalore was the first project that CPS completed successfully. It was also a pioneering project since it was the first in India to install an automated parking. Before the company's entry into the parking industry, the parking industry in India was mostly a cornucopia of unorganized players, which had not changed much even in 2013. After successful implementation of parking solution at various malls in Bangalore, CPS expanded their operation to other cities such as New Delhi and Pune. CPS was the first mover in a largely untapped but exponentially growing industry. However, many companies entered the market of organized parking solutions generating competition in the market.

Another very typical feature about the industry is that since the technology needed for the industry was fairly new, the entry barrier for new players was also very high. When asked, Poornima proudly stated:

Not only are we pioneers in this industry within India, what sets us apart is also our expertise in managing the whole end-to-end life cycle of the parking lot. We offer our customers a fully customized solution which includes not only the equipment but also a professional management staff and services to support on that.

As a strategy, CPS entered into tie-ups with various international players to suit local demands. In this way, it was able to offer customized state-of-the-art products and services to their customers. In order to stay ahead of the competition, it always believed in planning a sustainable strategy. In 2013, CPS controlled about 65% of the overall organized parking lot industry, and managed 65,000 parking bays across 32 cities.

### **The new direction**

In 2007–2008, Poornima's initiative led to data centralization resulting in all locations being connected to a single data center in one location instead of various data centers spread across different

**Continued...**

locations. This way, the management had access to live data across all the centers. However, the data that was captured was not fully utilized to reach business decisions.

Poornima said:

We are now growing very fast but if we have to grow faster, then we have to stop taking our decisions just by gut feeling and adopt a more analytical approach.

This thought found agreement with all members of the company's management. The company administration increasingly believed that with time, the need to understand the business increased. The company had to not only understand its immediate customers' requirements but also understand and predict the end customers' needs and behaviour. CPS divided its business into three main verticals, namely, airports, hospitals, and retail outlets. For each vertical, the critical factor which governed customer behaviour was different. CPS wanted to understand these factors in order to make better decisions such as reaching an appropriate pricing strategy for each vertical or optimizing the workforce management.

### Current Pricing Structure

Although CPS already had a pricing model (shown in **Exhibit 2**), this model mainly copied the pricing structure of other malls in similar regions. This pricing structure was only applicable to vehicles on regular tariff. The other pricing structure was meant for vehicles possessing pass cards. These pass cards were given to a few customers who were allotted dedicated parking spots. They could be shopkeepers, office staff, or a few privileged customers. These pass keepers did not necessarily make substantial contribution to the overall revenue of the company. These were more of additional or courtesy privileges that the company offered as mostly goodwill gestures and only a small percentage of parking slots were dedicated to pass holders.

### PROBLEM AT HAND

The company was struggling with operational issues not only at Howards End, but across many sites in the metropolitan cities where problems such as demand supply gap and waiting time at the entry and exits were becoming a daily occurrence. CPS management agreed that predicting the demand for number of bays on any given day and the length of stay was important for effective management of the parking lots. If the company could predict the spike in traffic, it would have two options: hire more manpower for those days or increase the workforce during specific intervals of time. But trained workforce was not only difficult to procure but also expensive. Even if it hired more people and trained them, the cost of both options was high and could bring its margins down.

According to the company management, if it had to hire more people, then would it have to increase the charges for those time periods? There was, of course, another reason for the company's management to think about modifying the pricing. A parking lot has a fixed capacity, and if the incoming traffic is more than the capacity of the parking lot, then once the parking lot is full, the shift manager would have no option but to ask incoming vehicles to return which would mean revenue

**Continued...**

leakage for the parking lot. This was also owing to the fact that while the capacity of the parking lot remained constant, the rate of incoming traffic was not always equal to the rate of outgoing traffic. Similar to the rate of incoming traffic, the ‘Length of Stay’ of an incoming vehicle could also vary depending upon the time of the day.

Therefore, a vehicle occupying the parking lot during those peak hours had to pay the opportunity cost of availing the services when the demand was very high. The administration knew that since Howards End was located in one of the most posh areas of Mumbai, price sensitivity of customers was relatively lower than other areas. The above reasons forced the company to re-think its pricing policy. This way, the company could manage peak demand with both manpower as well as pricing management, thereby, keeping its customer satisfaction high while also improving its overall operating profit.

Poornima knew that with the availability of transactional data solution to operational problems was more accessible at present than it would have been in the past. Owing to the improvement in technology, much transactional data was available. Although possessing the data was a crucial requirement, the next step was to use it to first perform the right kind of analysis and then draw conclusions and eventually take decisions to counter such operational issues. In the instance of Howards End, the company had managed to capture 3 years’ transactional data. The data was provided in a very simple format. It contained nine fields as shown in Table 1. Sample data is shown in **Exhibit 3**.

In order to investigate the data, analysts collected a sample of 5,000 data points for weekdays and 4995 data points for weekends, randomly selected from the population of the data available<sup>4</sup>. Based on this sample data, the average length of stay of the vehicles was determined. What came across as the first observation was that the length of stay of vehicles varied greatly. The graph shown in **Exhibit 4** plots the number of vehicles represented by individual entries against time spent by each vehicle in the parking lot.

As shown in **Exhibit 2**, the tariff during a weekday was INR 30/- for the first 3 hours and then INR 10/- for every incremental hour. During a weekend, the tariff was INR 50/- for the first 3 hours and then INR 20/- for every incremental hour. Poornima wanted the analysts to statistically determine if 3 hours was the correct time interval after which they could introduce an incremental tariff. Also, Poornima was

**TABLE 1** Data description

Data Code	Description
Vehicle	Type of vehicle (2W for 2-wheeler and 4W for 4-wheeler)
Date In	Date of entry of the vehicle
Time In	Time of entry of the vehicle
Date Out	Date of exit of the vehicle
Time Out	Time of exit of the vehicle
Amount	Parking fee paid
Length of Stay	Length of stay in the parking lot
Mode	Category of pay (ticket for regular user and pass for pass holders)
Weekday	Day of the week (Sunday, Monday, ..., Saturday)

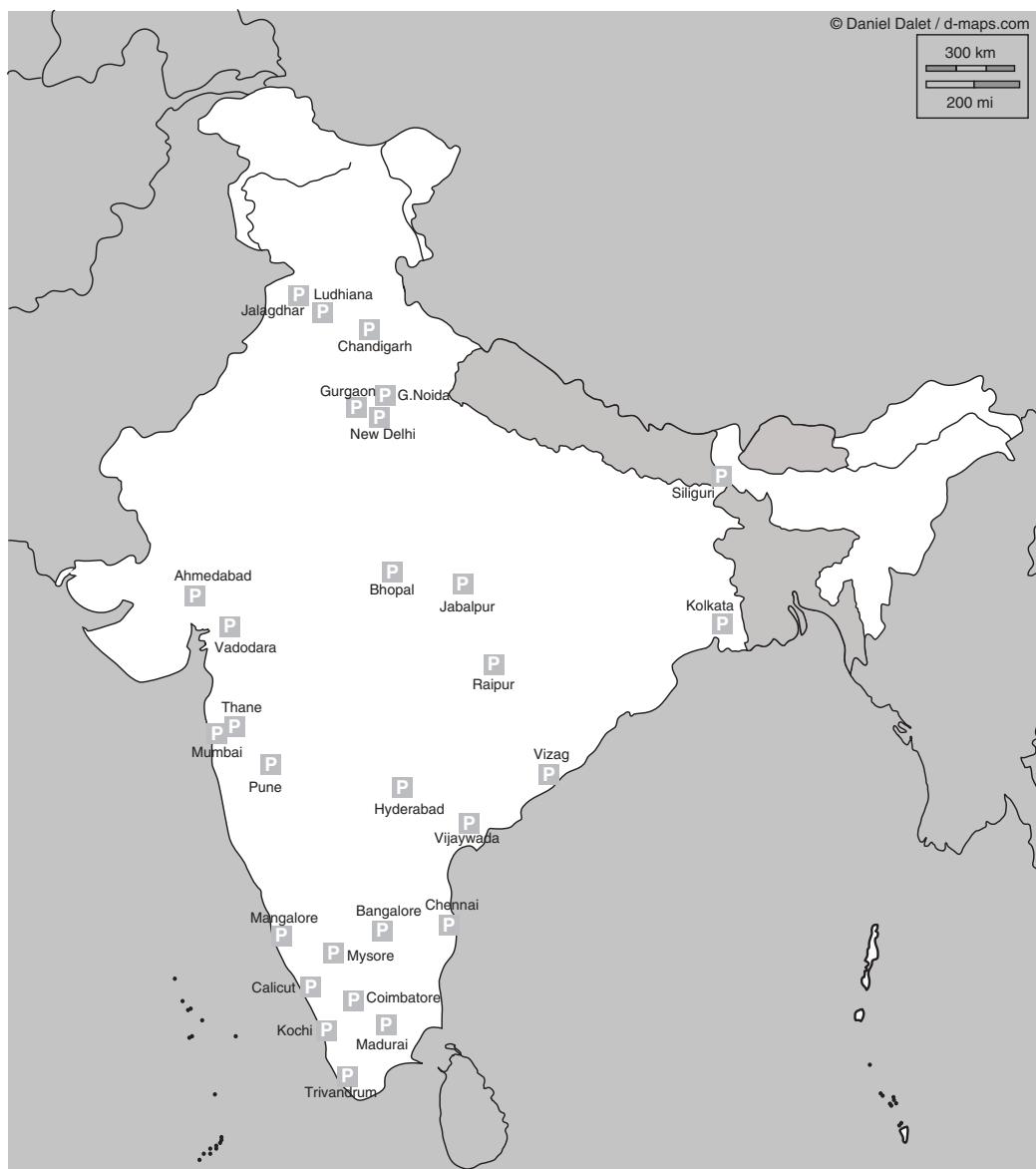
<sup>4</sup> Data set is provided in the spreadsheet supplement IMB453CPS.xls

Case Study

Continued...

interested in knowing whether they should use differential pricing across different time periods such as morning, afternoon, and evening. Poornima also questioned whether the average length of stay of vehicles really differed for cars coming in the morning, afternoon, and evening.

Also, according to **Exhibit 2**, the difference in the parking charges for the first 3 hours during a weekday versus a weekend was INR 20/- and the difference between the add-on charges or every



**EXHIBIT 1** CPS operating sites in India. Source: d-maps.com.

**Continued...**

**EXHIBIT 2** Pricing chart

S. No.	Description	Day of the Week	Parking Duration	Charges (in INR)		
1	4-wheeler	Monday to Friday	0 to 3 hours	30/-		
			Every Additional Hour	10/-		
	4-wheeler	Saturday, Sunday, Holidays	0 to 3 hours	50/-		
			Every Additional Hour	20/-		
2	Valet Parking	Monday to Friday	0 to 3 hours	100/-		
			Every Additional Hour	10/-		
		Saturday, Sunday, Holidays	0 to 3 hours	100/-		
			Every Additional Hour	20/-		
3	Loss of Ticket				300/-	
4	Government Vehicles				No Charge	

SOURCE: Central parking services.

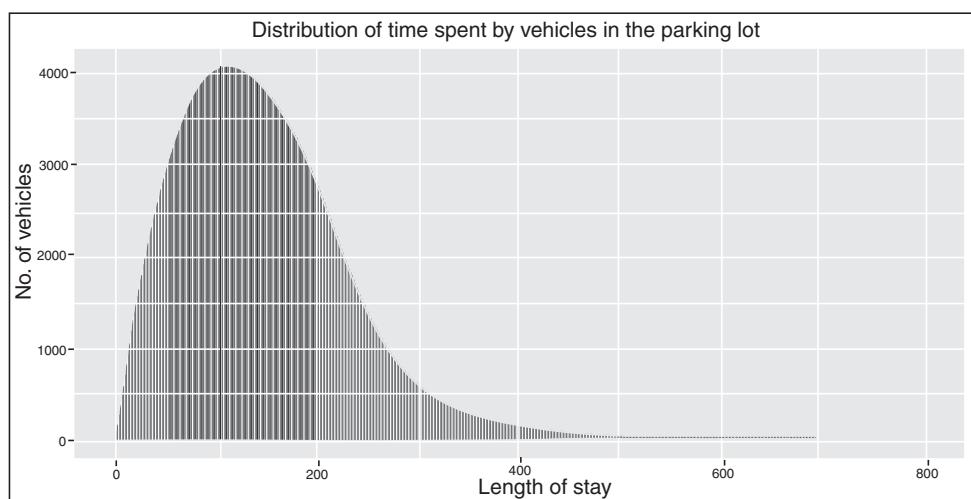
**EXHIBIT 3** Sample transaction data

Vehicle	Date In	Time In	Date Out	Time Out	Amount	Length of Stay (in minutes)	Mode	Weekday
4W	17-Sep-11	17:30:00	17-Sep-11	18:07:00	50	37	Ticket	Saturday
4W	12-Dec-10	19:42:00	12-Dec-10	21:31:00	50	109	Ticket	Sunday
4W	8-Nov-09	15:13:00	8-Nov-09	19:29:00	90	256	Ticket	Sunday
4W	16-Jan-11	15:11:00	16-Jan-11	16:09:00	50	58	Ticket	Sunday
4W	28-Nov-10	20:02:00	28-Nov-10	21:41:00	50	99	Ticket	Sunday
4W	16-Jul-11	14:53:00	16-Jul-11	16:34:00	10	101	Ticket	Saturday
4W	17-Jul-11	13:00:00	17-Jul-11	20:21:00	15	441	Ticket	Sunday
4W	20-Feb-11	11:56:00	20-Feb-11	12:28:00	50	32	Ticket	Sunday
4W	14-Feb-10	9:54:00	14-Feb-10	10:18:00	50	24	Ticket	Sunday
4W	28-Jan-12	15:04:00	28-Jan-12	19:57:00	90	293	Ticket	Saturday
4W	12-Jul-09	14:05:00	12-Jul-09	14:40:00	50	35	Ticket	Sunday
4W	30-May-10	18:52:00	30-May-10	20:27:00	50	95	Ticket	Sunday

SOURCE: Central parking services.

Case Study

**Continued...**



**EXHIBIT 4** Distribution of time spent by vehicles in parking lot on a certain weekend. Source: central parking services.

additional hour after the initial 3 hours was INR 10/-. What the company needed to decipher was whether this differentiated pricing policy for weekdays and weekends a suitable strategy or would it make no difference to the revenue earned if it kept a uniform tariff for all days of the week.

CPS also noticed that during the weekend, the arrival rate was significantly high during the matinee time (2 pm to 6 pm) and was sometimes nearly twice that of arrival rate at other times of the day. This observation also makes one to think whether it will be more advisable to deploy more manpower to manage the incoming traffic more efficiently and effectively during those hours. If CPS could determine the probability of the incoming traffic being significantly high, then it would be easier to plan manpower deployment, so that it could provide appropriate and desired service to customers during those hours, thereby, keeping customer satisfaction high.

#### CASE QUESTIONS

1. Explore the data (file name: CPS CASE DATA.xls) for weekday and weekend. What inferences can you make based on descriptive statistics?
2. Test whether the random variable 'length of stay' for weekend and weekday follows normal distribution.
3. Between normal and Weibull distribution, which distribution should be used to represent the random variable 'length of stay' during weekend?
4. CPS charges for weekends are more than weekdays. One of the reasons for higher parking fee during weekends is that the customers tend to stay for longer duration resulting in non-availability of parking lots. Is there an evidence to support that the customers stay for longer period during weekends compared to weekdays?
5. If we divide the day into three time periods, viz., 10 am to 2 pm as Morning, 2 pm to 6 pm as Afternoon, and 6 pm to 10 pm as Evening, then determine if the average length of stay really varies between these three time periods.

**Continued...**

6. CPS charges its customers INR 30 for first 3 hours and INR 10 for every additional hour on weekdays. What will be the financial impact to CPS if they charge INR 20 for the first 2 hours and INR 10 for every additional hour? What assumptions are made in this analysis? (\$1 = INR 62, in November 2013)
7. The average arrival rate of cars during time periods for Sunday and Monday are shown in the following table. How many additional workers are required on Sundays compared to Mondays if CPS employs one additional worker for every 30 vehicles entering the parking lot?

**Arrival Rate on Monday and Sunday**

Time Interval	Monday	Sunday
10 am – 11 am	30	42
11 am – 12 pm	74	113
12 pm – 1 pm	120	211
1 pm – 2 pm	147	271
2 pm – 3 pm	155	286
3 pm – 4 pm	172	278
4 pm – 5 pm	173	286
5 pm – 6 pm	163	279
6 pm – 7 pm	153	292
7 pm – 8 pm	162	285
8 pm – 9 pm	134	204
9 pm – 10 pm	74	101

8. According to Poornima, the average occupancy time on a weekday is utmost 2 hours, but the analyst's calculation on data showed otherwise. She was still not assured and asked analysts to check again. Analysts thought of conducting a statistical test over Poornima's data. Please help the analyst in coming up with the test at 0.05 level of significance. Assume sample's variance to be same as the population variance.

**REFERENCES**

1. Anon (2009), "Uganda blackout fuel baby boom". BBC News 12 March 2009, available at <http://news.bbc.co.uk/2/hi/africa/7939534.stm>. Accessed on 1 May 2017.
2. Anon (2015), "There really is such a thing as wedded bliss: Married couple are more happier than singles says a new study following benefits of matrimony", *Mail Online*, 12 January 2015. Available at <http://www.dailymail.co.uk/news/article-2904986/There-really-thing-wedded-bliss-Married-couples-happier-singles-says-new-study-following-benefits-matrimony.html> accessed on 30 March 2017.
3. Anon (2017), "World Ranking of Countries by their Average IQ", available at <https://iq-research.info/en/page/average-iq-by-country> accessed on 27 March 2017.
4. Cohen J (1977), "Statistical Power Analysis for the Behavioural Sciences", Academic Press, California.
5. Elston R. and Johnson W (2008), "Basic Biostatistics for Geneticists and Epidemiologists: A Practical Approach", John Wiley and Sons, New York.

6. Fetzer T, Pardo O and Shanghavi A, (2013), "An Urban Legend? Power Rationing, Fertility and its Effect on Mothers", CEP Discussion Paper, Centre for Economic Performance, London School of Economics.
7. Fisher R A (1956), "*Statistical Methods and Scientific Inference*", Hafner Publishing Company, New York.
8. Freier A (2016), "Women spend more Time using their Smartphones than Men", *Business of Apps*, June 6 2016, available at <http://www.businessofapps.com/women-spend-more-time-using-their-smartphones-than-men/> accessed on 30 March 2017.
9. Hubbard R, Bayarri M J, Berk K N and Carlton M A, (2003), "Confusion over Measures of Evidence (p's) versus Errors ( $\alpha$ 's) in Classical Statistical Testing", *The American Statistician*, **57**(3), 171–182.
10. Izenman A J and Zabell S L (1981), "Babies and blackout: The genesis of misconception", *Social Science Research*, **10**(3), 282–299.
11. Mann H B and Wald A (1942), "On the choice of the number of class intervals in the Application of the Chi Square test", *The Annals of Mathematical Statistics*, **13**(3), 306–317.
12. Miller A S and Kanazawa S (2007), "Why Beautiful People have more Daughters", *A Perigee Book*, New York, 2007.
13. Sawyer A G and Ball A D (1981), "Statistical Power and Effect Size in Marketing Research", *Journal of Marketing Research*, **18**(3), 275–290.
14. Siegel E (2016), "*Predictive Analytics: The Power to Predict who will click, buy, lie or Die*", Wiley, New York, 2016.
15. Sturges H A (1926), "The Choice of a Class Interval", *Journal of the American Statistical Association*, **21**(153), 65–66.
16. Tukey J W (1977), "*Exploratory Data Analysis*", Addison-Wesley Publishing Co, Reading, MA.



# 7

# Analysis of Variance

“Analysis of variance is not a mathematical theorem, but rather a convenient method of arranging the arithmetic”.

— Ronald Fisher

## LEARNING OBJECTIVES

- LO 7-1** Understand the need for Analysis of Variance (ANOVA).
- LO 7-2** Understand the difference between two-sample  $t$ -test for mean and ANOVA.
- LO 7-3** Understand one-way ANOVA and calculation of  $F$ -statistic.
- LO 7-4** Understand computation within the group variation, between the group variation and  $F$ -statistic.
- LO 7-5** Learn to conduct a two-way ANOVA and the computations involved in conducting a two-way ANOVA.

## ANALYSIS OF VARIANCE

In many situations we may have conduct a hypothesis test to compare mean values simultaneously for more than two groups (samples) created using a factor (or factors). For example, a marketer may like to understand the impact of three different discount values (such as 0%, 10%, and 20% discount) on the average sales. When we have to compare the impact of a factor on mean on more than two groups (created by different levels of the factor) simultaneously, the hypothesis tests such as two-sample  $t$ -tests discussed in Chapter 6 are not ideal approach since they can result in incorrect Type I and Type II errors. We use the Analysis of Variance (ANOVA) to understand the differences in population means among more than two populations.



*The objective of ANOVA is to check simultaneously whether population mean from more than two populations are different.*

## 7.1 | INTRODUCTION TO ANALYSIS OF VARIANCE (ANOVA)

Consider a retail store which would like to study the impact of different levels of price discounts (factor) on sales (outcome variable) of a specific product or brand. Price discount can range from 0% to 100% (theoretically). For easier understanding, assume that the levels of discounts are 0%, 10%, and 20%.

The marketing manager would like to understand whether the variable 'price discount' has any significant impact on the average sales quantity. Such studies are called **single factor experimental studies**. Different discount rates correspond to different levels of the factor (different levels of price discount) and different levels (such as 0%, 10% and 20%) are assigned randomly to different units. In the case of price discounts, units refer to different days chosen randomly since the quantity of sales may also depend on the day of the week. It is possible that the weekend sales quantity may be higher than the sales quantity on weekdays. In many cases, we may deal with observational studies in which we observe the impact of a factor on a variable, for example, impact of specialization in MBA such as analytics, finance, marketing, etc. on the graduating income of the graduates. Here the units are not subjected to experiment, which is probably not under the control of the researcher. To understand whether the effect (different levels of a factor) has any statistical significance on the population parameter, we compare two models as described below:

- 1. Means Model:** It is given by

$$Y_{ij} = \mu + \varepsilon_{ij} \quad (7.1)$$

where  $Y_{ij}$  is the value of the outcome variable of  $j^{\text{th}}$  observation for  $i^{\text{th}}$  factor level,  $\mu$  is the overall mean value of all observations,  $\varepsilon_{ij}$  is the error assumed to be a normal distribution with mean 0 and standard deviation  $\sigma$ . Model defined in Eq. (7.1) is often called the reduced model, in which the mean  $\mu$  is common for all levels of the factor.

Since we assume that the error  $\varepsilon_{ij}$  is normal distribution with mean 0 and standard deviation  $\sigma$ , the outcome variable  $Y_{ij}$  is a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ .

- 2. Factor Effect Model:** It is given by

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad (7.2)$$

In Eq. (7.2),  $\mu$  is the overall mean and  $\tau_i$  is the effect of factor  $i$  (or factor effect).  $\tau_i$  is the difference between overall mean and the factor level mean. Our interest in this case would be to check whether the values of  $\tau_i$  are different from zero. The model in Eq. (7.2) is called full model. The reduced model in Eq. (7.1) is a special case of model defined in Eq. (7.2) in which  $\tau_i$  is zero for all  $i$ .




---

A non-zero  $\tau_i$  value in Eq. (7.2) implies that the factor has influence on the value of the outcome variable  $Y_{ij}$ .




---

In ANOVA, our objective is to verify whether the variation due to treatment is different from the variation due to randomness.

## 7.2 | MULTIPLE $t$ -TESTS FOR COMPARING SEVERAL MEANS

Continuing with the example from Section 7.1, if we had only two values for ‘price discount’, then we could have used the two-sample  $t$ -test to check whether there is a statistically significant relationship between price discount and average sales quantity. When we have more than two levels of discounts, one option is to compare the population parameters two at a time (two discount values). For example, we can compare the following three cases using two-sample  $t$ -test:

1. Test between 0% and 10%
2. Test between 0% and 20%
3. Test between 10% and 20%

However, when we want to test the hypothesis simultaneously, the Type I and Type II errors will not be same if we conduct the three different tests listed above. For example, assume that the mean sale (population mean) at 0%, 10%, and 20% discount is  $\mu_0$ ,  $\mu_{10}$ , and  $\mu_{20}$ , respectively. Consider the following three two-sample  $t$ -tests shown in Table 7.1.

Let

$$P(A) = P(\text{Retain } H_0 \text{ in test A} | H_0 \text{ in test A is true})$$

$$P(B) = P(\text{Retain } H_0 \text{ in test B} | H_0 \text{ in test B is true})$$

$$P(C) = P(\text{Retain } H_0 \text{ in test C} | H_0 \text{ in test C is true})$$

Note that values of  $P(A) = P(B) = P(C) = 1 - \alpha = 1 - 0.05 = 0.95$

The conditional probability of simultaneously retaining all 3 null hypotheses when they are true is  $P(A \cap B \cap C) = 0.8573$ . Now consider the following null hypothesis:

$$H_0: \mu_0 = \mu_{10} = \mu_{20} \quad (7.3)$$

If we retain the null hypothesis based on the three individual  $t$ -tests, then the significance or Type I error is not  $\alpha$ -value, but much higher than  $\alpha$  (Lunney, 1969; Siegel, 1990). For the case discussed above, if we retain the null hypothesis based on 3 individual tests, then the Type I error is  $1 - 0.8573 = 0.1426$ . That is, when more than 2 groups are involved, checking the population parameter values simultaneously using  $t$ -tests is inappropriate since the Type I and Type II errors will be estimated incorrectly. For this reason, we use analysis of variance (ANOVA) whenever we need to compare 3 or more groups for population parameter values simultaneously.

**TABLE 7.1** Three different two-population  $t$ -tests

Test	Null Hypothesis	Alternative Hypothesis	Significance ( $\alpha$ )
A	$H_0: \mu_0 = \mu_{10}$	$H_A: \mu_0 \neq \mu_{10}$	$\alpha = 0.05$
B	$H_0: \mu_0 = \mu_{20}$	$H_A: \mu_0 \neq \mu_{20}$	$\alpha = 0.05$
C	$H_0: \mu_{10} = \mu_{20}$	$H_A: \mu_{10} \neq \mu_{20}$	$\alpha = 0.05$

## 7.3 | ONE-WAY ANALYSIS OF VARIANCE (ANOVA)

One-way ANOVA is appropriate under the following conditions:

1. We would like to study the impact of a single treatment (also known as factor) at different levels (thus forming different groups) on a continuous response variable (or outcome variable). For the example discussed in Section 7.1, the variable ‘price discount’ is the treatment (or factor) and 0%, 10%, and 20% price discounts are the different levels (3 levels in this case); different levels of discount are likely to have varying impact on the sales of the product, where sales is the outcome variable. We would like to understand the impact of different levels of price discount on the response variable, sales. The term ‘treatment’ is used since one of the initial applications of ANOVA was to find the impact of different fertilizer treatments on agricultural yield as studied by British statistician R A Fisher (1934).
2. In each group, the population response variable follows a normal distribution and the sample subjects are chosen using random sampling.
3. The population variances for different groups are assumed to be same. That is, variability in the response variable values within different groups is same.

Although conditions 2 and 3 are necessary for one-way ANOVA, the model is robust and minor violations of the assumptions may not result in incorrect decision about the null hypothesis. However, we need to check whether conditions 2 and 3 are met to ensure validity of ANOVA as a good practice. Normality assumption can be checked either using P-P plot (probability-probability plot) or using goodness of fit tests such as chi-square goodness of fit test. The equality of variance can be checked through the hypothesis test for equal variance discussed in Chapter 6.

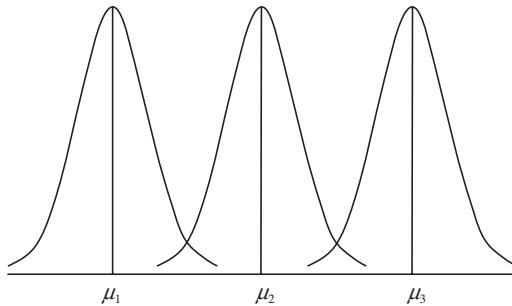
### 7.3.1 | Setting up an Analysis of Variance

Assume that we would like to study the impact of a factor (such as discount) with  $k$  levels on a continuous variable (such as sales quantity). Then the null and alternative hypotheses for one-way ANOVA are given by

$$\begin{aligned} H_0: \mu_1 &= \mu_2 = \mu_3 = \dots = \mu_k \\ H_A: \text{Not all } \mu \text{ values are equal} \end{aligned}$$

Note that the alternative hypothesis, ‘not all  $\mu$  values are equal’, implies that some of them could be equal. The null hypothesis is equivalent to stating that the factor effects  $\tau_1, \tau_2, \dots, \tau_k$  defined in Eq. (7.2) are zero. The hypothesis test can be visualized as shown in Figure 7.1. Different values of mean ( $\mu_1, \mu_2$ , and  $\mu_3$ ) imply statistically significant impact of factor levels on the response variable. We expect the group means ( $\mu_1, \mu_2$ , and  $\mu_3$ ) to be closer to one another if the factor levels do not have any impact.

If the mean values of different groups are not equal, then the variation of cases within the group will be much smaller compared to variations between groups. Assume that we are interested in analysing single factor effect with  $k$  levels; thus we will have  $k$  groups.



**FIGURE 7.1** Comparing three means ( $\mu_1$ ,  $\mu_2$ , and  $\mu_3$ ).

Let

$k$  = Number of groups (or samples)

$n_i$  = Number of observations in group  $i$  ( $i = 1, 2, \dots, k$ )

$n$  = Total number of observations  $\left(= \sum_{i=1}^k n_i\right)$

$Y_{ij}$  = Observation  $j$  in group  $i$

$\mu_i$  = Mean of group  $i$   $= \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$

$\mu$  = Overall mean  $= \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij}$

To arrive at the statistic, we calculate the following measures, which are variations within group and between groups:

1. **Sum of Squares of Total Variation (SST):** Total variation is the sum of squares variation of all values of response variable ( $Y_{ij}$ ) from the overall mean ( $\mu$ ) and is given by

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \mu)^2 \quad (7.4)$$

The degrees of freedom for SST is  $(n - 1)$  since only the value of  $\mu$  is estimated from  $n$  observations and thus only one degree of freedom is lost. Mean Square Total (MST) variation is given by

$$MST = \frac{SST}{n-1} \quad (7.5)$$

2. **Sum of Squares of Between (SSB) Group Variation:** Sum of squares of between group variation is the sum of squares variation between the group mean ( $\mu_i$ ) and the overall mean ( $\mu$ ) of the data and is given by

$$SSB = \sum_{i=1}^k n_i \times (\mu_i - \mu)^2 \quad (7.6)$$

The degrees of freedom for  $SSB$  is  $(k - 1)$ . Since the overall mean  $\mu$  is estimated from the data, one degree of freedom is lost. Mean square between variation ( $MSB$ ) is given by

$$MSB = \frac{SSB}{k-1} \quad (7.7)$$

- 3. Sum of Squares of Within (SSW) Group Variation:** Sum of squares of within the group variation is the sum of squares variation of all observations ( $Y_{ij}$ ) from that group mean ( $\mu_i$ ) and is given by

$$SSW = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2 \quad (7.8)$$

The degrees of freedom for  $SSW$  is  $(n - k)$ . Here  $k$  degrees of freedom are lost since we estimate  $k$  group means ( $\mu_i$ ). The mean square of variation within the group is

$$MSW = \frac{SSW}{n - k} \quad (7.9)$$

We can prove algebraically

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \mu)^2 = \sum_{i=1}^k n_i \times (\mu_i - \mu)^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2 \quad (7.10)$$

That is

$$SST = SSB + SSW \quad (7.11)$$

### 7.3.2 | Cochran's Theorem

According to Cochran's theorem (Kutner *et al.*, 2013, page 70):

"If  $Y_1, Y_2, \dots, Y_n$  are drawn from a normal distribution with mean  $\mu$  and standard deviation  $\sigma$  and sum of squares of total variation [Eq. (7.11)] is decomposed into  $k$  sum of squares ( $SS_r$ ) with degrees of freedom  $df_r$ , then the ratio  $(SS_r/\sigma^2)$  are independent  $\chi^2$  variables with  $df_r$  degrees of freedom if  $\sum_{r=1}^k df_r = n - 1$ ."

Note that, in Eq. (7.11) the  $SST$  is decomposed into two sums of squares ( $SSB$  and  $SSW$ ) and thus,  $SSB/\sigma^2$  and  $SSW/\sigma^2$  are chi-square variables.

### 7.3.3 | THE F-TEST

If the null hypothesis is true, then there will be no difference in the mean values which will result in no difference between  $MSB$  and  $MSW$ . Alternatively, if the means are different, then  $MSB$  will be larger than  $MSW$ . That is the ratio  $MSB/MSW$  will be close to 1 if there is no difference between the mean values and will be larger than 1 if the means are different. Following Cochran's theorem (Kirk, 1995)  $MSB/MSW$  is a ratio of two chi-square variate which is an  $F$ -distribution. Thus the statistic for testing the null hypothesis is

$$F = \frac{SSB / (k - 1)}{SSW / (n - k)} = \frac{MSB}{MSW} \quad (7.12)$$

Note that the test statistic is a one-tailed test (right tailed) since we are interested in finding whether the variation between groups is greater than variation within the groups. Although we are checking whether

the means are equal in the null hypothesis, the actual testing is carried out by checking whether the variation between groups is higher than within the groups, thus it is a one-tailed (right-tailed) test. It is important to note that rejecting the null hypothesis will not tell us exactly which means differ from each other, but it will only indicate that there is a difference in at least one of the group means. We may have to conduct two-sample  $t$ -tests to find which mean values are different.

**EXAMPLE 7.1**

Ms Rachael Khanna the brand manager of ENZO detergent powder at the ‘one stop’ retail was interested in understanding whether the price discounts has any impact on the sales quantity of ENZO. To test whether the price discounts had any impact, price discounts of 0% (no discount), 10% and 20% were given on randomly selected days. The quantity (in kilograms) of ENZO sold in a day under different discount levels is shown in Table 7.2. Conduct a one-way ANOVA to check whether discount had any significant impact on the average sales quantity at  $\alpha=0.05$ .

**TABLE 7.2** Sales of ENZO at different price discounts

No Discount (0% discount)									
39	32	25	25	37	28	26	26	40	29
37	34	28	36	38	38	34	31	39	36
34	25	33	26	33	26	26	27	32	40
10% Discount									
34	41	45	39	38	33	35	41	47	34
47	44	46	38	42	33	37	45	38	44
38	35	34	34	37	39	34	34	36	41
20% Discount									
42	43	44	46	41	52	43	42	50	41
41	47	55	55	47	48	41	42	45	48
40	50	52	43	47	55	49	46	55	42

**Solution:**

In this case, the number of groups  $k = 3$ ;  $n_1 = n_2 = n_3 = 10$ ;  $\mu_1 = 39.05$ ,  $\mu_2 = 38.77$ ,  $\mu_3 = 46.4$ ; and  $\mu = 39.05$ .

The sum of squares of between groups variation (SSB) is given by

$$\begin{aligned} SSB &= \sum_{i=1}^k n_i \times (\mu_i - \mu)^2 = 10 \times [(39.05 - 39.05)^2 + (38.77 - 39.05)^2 \\ &\quad + (46.4 - 39.05)^2] = 3114.156 \end{aligned}$$

So

$$MSB = \frac{SSB}{k-1} = \frac{3114.156}{2} = 1557.078$$

The sum of squares of within the group variation is given by

$$\begin{aligned}SSW &= \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2 = \sum_{j=1}^{30} (Y_{1j} - 32)^2 + \sum_{j=1}^{30} (Y_{2j} - 38.77)^2 \\&\quad + \sum_{j=1}^{30} (Y_{3j} - 46.4)^2 = 2056.567 \\MSW &= \frac{SSW}{n-k} = \frac{2056.567}{90-3} = 23.63\end{aligned}$$

The  $F$ -statistic value is

$$F_{2,87} = \frac{MSB}{MSW} = \frac{1557.078}{23.6387} = 65.86$$

The critical  $F$ -value with degrees of freedom (2, 87) for  $\alpha=0.05$  is 3.101 [Excel function FINV(0.05, 2, 87) or F.INV.RT(0.05, 2, 87)]. The  $p$ -value for  $F_{2,87}=65.86$  is  $3.82 \times 10^{-18}$  [using Excel function FDIST(65.86, 2, 87) or F.DIST.RT(65.86, 2, 87)]. Since the calculated  $F$ -statistic is much higher than the critical  $F$ -value, we reject the null hypothesis and conclude that the mean sales quantity values under different discounts are different. The Excel output of ANOVA is shown in Table 7.3.

**TABLE 7.3** One-way ANOVA excel output for Example 7.1

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
No Discount	30	960	32	27.17241		
10% Discount	30	1163	38.76667	20.46092		
20% Discount	30	1392	46.4	23.28276		
ANOVA						
Source of Variation	SS	df	MS	F	p-value	F crit
Between Groups	3114.15556	2	1557.078	65.86986	3.82E-18	3.101296
Within Groups	2056.56667	87	23.6387			
Total	5170.72222	89				

Example 7.1 is an experimental study in which the marketer was trying to study the impact of discounts on sales. Example 7.2 is an observational study in which we understand the impact of different sectors on stock returns.

**EXAMPLE 7.2**

Share Raja Khan (SRK) is a top stockbroker and believes that the average annual stock return depends on the industrial sector. To validate his belief, SRK collected annual return of shares from three different industrial sectors – consumer goods, services, and industrial goods. The annual return of shares in 2015–2016 for different sectors is shown in Table 7.4.

**TABLE 7.4** Annual return of stocks under different industrial sector

Annual return on 30 consumer goods stocks									
6.32%	14.73%	11.95%	12.36%	10.28%	3.81%	10.15%	11.06%	6.29%	5.15%
8.44%	14.28%	8.89%	5.98%	6.96%	11.62%	5.22%	5.34%	5.93%	7.10%
10.91%	8.20%	10.19%	9.04%	8.61%	9.39%	2.63%	2.77%	4.76%	9.60%
Annual return on 30 services stocks									
13.70%	3.58%	1.36%	17.41%	10.01%	10.88%	15.63%	-0.04%	10.32%	7.40%
11.48%	9.71%	11.19%	8.21%	1.64%	1.45%	10.12%	13.85%	-10.27%	5.26%
12.05%	4.47%	8.71%	5.59%	10.02%	7.65%	10.03%	7.87%	6.59%	13.60%
Annual return on 30 industrial goods stocks									
6.74%	7.11%	5.69%	2.48%	5.42%	8.00%	2.55%	8.34%	4.99%	3.39%
8.73%	13.85%	5.29%	9.06%	2.84%	5.82%	7.66%	4.12%	9.10%	8.76%
10.77%	1.48%	4.71%	10.66%	0.44%	2.94%	6.55%	2.84%	3.90%	7.28%

**Solution:** In this case, the number of cases  $k = 3$ ;  $n_1 = n_2 = n_3 = 30$ ;  $\mu_1 = 0.082$ ,  $\mu_2 = 0.079$ ,  $\mu_3 = 0.0605$ ; and  $\mu = 0.0743$ .

The sum of squares of between groups (SSB) variation is given by

$$\begin{aligned} SSB &= \sum_{i=1}^k n_i \times (\mu_i - \mu)^2 = 30 \times [(0.082 - 0.0743)^2 + (0.079 - 0.0743)^2 \\ &\quad + (0.0605 - 0.0743)^2] = 0.0087 \end{aligned}$$

Therefore

$$MSB = \frac{SSB}{k-1} = \frac{0.0087}{2} = 0.0043$$

The sum of squares of within the group variation is given by

$$\begin{aligned} SSW &= \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2 = \sum_{j=1}^{30} (Y_{1j} - 0.082)^2 + \sum_{j=1}^{30} (Y_{2j} - 0.079)^2 \\ &\quad + \sum_{j=1}^{30} (Y_{3j} - 0.0605)^2 = 0.1463 \end{aligned}$$

So

$$MSW = \frac{SSW}{n - k} = \frac{0.1463}{90 - 3} = 0.0016$$

The  $F$ -statistic value is

$$F_{2,87} = \frac{MSB}{MSW} = \frac{0.0043}{0.0016} = 2.592$$

The critical  $F$ -value with degrees of freedom (2, 87) for  $\alpha = 0.05$  is 3.101 [Excel function FINV(0.05, 2, 87) or F.INV.RT(0.05, 2, 87)]. The  $P$ -value for  $F_{2,87} = 2.592$  is 0.0805 [using Excel function FDIST(2.592, 2, 87) or F.DIST.RT(2.592, 2, 87)]. Since the calculated  $F$ -statistic is less than the critical  $F$ -value, we retain the null hypothesis and conclude that the average annual returns under industrial sectors consumer goods, services, and industrial goods are not different (Figure 7.2 shows the  $F$ -critical value and  $F$ -statistic value for an  $F$ -distribution with degrees of freedom 2 and 87 for numerator and denominator, respectively). The Excel output of ANOVA is shown in Table 7.5.

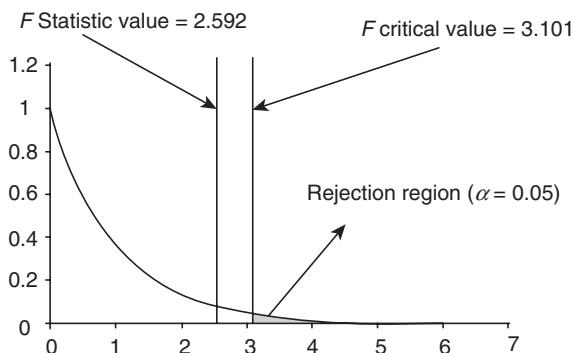


FIGURE 7.2  $F$ -distribution with critical value for Example 7.2.

TABLE 7.5 Microsoft excel ANOVA table for Example 7.2

ANOVA: Single Factor						
SUMMARY						
<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>		
Consumer Goods	30	2.4796	0.082653	0.00101		
Services	30	2.3947	0.079823	0.003073		
Industrial Goods	30	1.8151	0.060503	0.000963		
ANOVA						
<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p-value</i>	<i>F critical</i>
Between Groups	0.008722	2	0.004361	2.59294	0.080572	3.101296
Within Groups	0.146317	87	0.001682			
Total	0.155039	89				

## 7.4 | TWO-WAY ANALYSIS OF VARIANCE (ANOVA)

The values of response variable may be influenced by several factors. For example, in addition to price discounts, location of the stores may also play an important role in the sales quantity. The discounts may not have much impact if the store is located near affluent community compared to stores located near non-affluent community. We would like to understand the impact of both factors (price discount and location) simultaneously on sales by trying to answer to the following questions:

1. Are there differences in the average sales quantity with different levels of price discounts?
2. Are there differences in the average sales quantity with respect to different locations?
3. Are there interactions between price discounts and location with respect to average sales quantity?

The two-way ANOVA model can be expressed as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha_i\beta_j + \varepsilon_{ijk} \quad (7.13)$$

where  $Y_{ijk}$  = Value of the  $k^{\text{th}}$  observation ( $k = 1, 2, \dots, c$ ) of the response variable at level  $i$  ( $i = 1, 2, \dots, a$ ) of factor A and level  $j$  ( $j = 1, 2, \dots, b$ ) of factor B.

$\mu$  = Overall mean value of the response variable  $Y_{ijk}$

$\alpha_i$  = Level (effect) of factor A ( $i = 1, 2, \dots, a$ )

$\beta_j$  = Level (effect) of factor B ( $j = 1, 2, \dots, b$ )

$\alpha_i\beta_j$  = Interaction of  $i^{\text{th}}$  level of factor A and  $j^{\text{th}}$  level of factor B

$\varepsilon_{ijk}$  = Error associated with  $k^{\text{th}}$  of observation at level  $i$  of factor A and level  $j$  of factor B.

The hypothesis tests associated with two-way ANOVA are as follows:

### 1. Test of Factor A Main Effects:

$$H_0: \alpha_i = 0 \text{ for all } i (i = 1, 2, \dots, a)$$

$$H_A: \text{Not all } \alpha_i \text{ are zero}$$

### 2. Test of Factor B Main Effects:

$$H_0: \beta_j = 0 \text{ for all } j (j = 1, 2, \dots, b)$$

$$H_A: \text{Not all } \beta_j \text{ are zero}$$

### 3. Test of Interaction Effects:

$$H_0: \alpha_i\beta_j = 0 \text{ for all } i (i = 1, 2, \dots, a) \text{ and } j (j = 1, 2, \dots, b)$$

$$H_A: \text{Not all } \alpha_i\beta_j \text{ are zero}$$

The sum of squares in the case of two-way ANOVA with equal sample sizes is given by (Fisher, 1934)

$$\text{SST} = \text{SSA} + \text{SSB} + \text{SSAB} + \text{SSE} \quad (7.14)$$

Various components in Eq. (7.14) are provided below:

**1. Sum of squares of total deviation (SST):**

$$SST = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c (Y_{ijk} - \mu)^2 \quad (7.15)$$

where  $c$  is the number of observations in each group and  $\mu$  is the overall mean.

**2. Sum of squares of deviation due to factor A (SSA):**

$$SSA = b \times c \times \sum_{i=1}^a (\mu_i - \mu)^2 \quad (7.16)$$

where  $\mu_i$  is the mean of all observations in level  $i$  of factor A and  $c$  is the number of observations in each group (assumed to be same for all groups).

**3. Sum of squares of deviation due to factor B (SSB):**

$$SSB = a \times c \times \sum_{j=1}^b (\mu_j - \mu)^2 \quad (7.17)$$

Here  $\mu_j$  is the mean of all observations in level  $j$  of factor B.

**4. Sum of squares of deviation due to interaction of factors A and B (SSAB):**

$$SSAB = c \times \sum_{i=1}^a \sum_{j=1}^b (\mu_{ij} - \mu_i - \mu_j + \mu)^2 \quad (7.18)$$

where  $\mu_{ij}$  is the average of  $i^{\text{th}}$  level of factor A and  $j^{\text{th}}$  level of factor B.

**5. Sum of squares of deviation within a group (SSW):**

$$SSW = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c (Y_{ijk} - \mu_{ij})^2 \quad (7.19)$$

Different factors, degrees of freedom, and  $F$ -statistic for two-way ANOVA with equal number of samples is given in Table 7.6.

**TABLE 7.6** Sum of squares of deviation for various effects and the corresponding  $F$ -statistic in a two-way ANOVA with equal sample size

Sum of Squares Variation	Degrees of Freedom	Mean Squared Variation	$F$ -Statistics
SSA	$a - 1$	$MSA = SSA/(a - 1)$	$F = MSA/MSW$
SSB	$b - 1$	$MSB = SSB/(b - 1)$	$F = MSB/MSW$
SSAB	$(a - 1)(b - 1)$	$MSAB = SSAB/(a - 1)(b - 1)$	$F = MSAB/MSW$
SSW	$ab(c - 1)$	$MSW = SSW/ab(c - 1)$	

**EXAMPLE 7.3**

Table 7.7 shows the sales quantity of detergents at different discount values and different locations collected over 20 days. Conduct a two-way ANOVA at  $\alpha = 0.05$  to test the effects of discounts and location on the sales.

**TABLE 7.7** Sales quantity at different locations under different discount rates

Location 1			Location 2		
Discount			Discount		
0%	10%	20%	0%	10%	20%
20	28	32	20	19	20
16	23	29	21	27	31
24	25	28	23	23	35
20	31	27	19	30	25
19	25	30	25	25	31
10	24	26	22	21	31
24	28	37	25	33	31
16	23	33	21	26	23
25	26	27	26	22	22
16	25	31	22	28	32
18	22	37	25	24	22
20	24	28	23	23	29
17	26	25	23	26	25
26	28	23	24	16	34
16	21	26	20	30	30
21	27	33	23	22	25
24	25	28	18	16	39
19	20	30	19	25	32
19	26	30	19	34	29
21	26	26	30	23	22

The two-way ANOVA with replication (since the data in Table 7.7 is repeated for locations) output from Microsoft Excel is shown in Table 7.8.

The two-way ANOVA with replication (since the data in Table 7.7 is repeated for locations) output from Microsoft Excel is shown in Table 7.8.

**TABLE 7.8** Two-way ANOVA with replication excel output

ANOVA						
Source of Variation	SS	df	MS	F	p-value	F crit
Sample (Location)	7.008333	1	7.008333	0.443898	0.506593	3.92433
Columns (Discount)	1240.317	2	620.1583	39.27997	1.06E-13	3.075853
Interaction	84.81667	2	42.40833	2.686085	0.07246	3.075853
Within	1799.85	114	15.78816			
Total	3131.992	119				

In Table 7.8, the sample stands for the row factor (which in this case is location), column stands for the column factor (discount in this case), and interaction stands for interaction effect (location  $\times$  discount). The *p*-value for locations (data in rows) is 0.5065, thus it is not statistically significant (we retain the null hypothesis that the locations have no statistically significant influence on sales), whereas for discount rates (data in column) the *p*-value is  $1.06 \times 10^{-13}$ , so we reject the null hypothesis (that is discount rate has influence on sales). The *p*-value for the interaction effect is 0.0724 and is not significant. That is, only the factor discount is statistically significant at  $\alpha=0.05$ .

## SUMMARY

1. Analysis of Variance (ANOVA) is a hypothesis testing procedure used for comparing means from several groups simultaneously.
2. In a one-way ANOVA, we test whether the mean values of an outcome variable for different levels of a factor are different. Using multiple two-sample *t*-tests to simultaneously test group means will result in incorrect estimation of Type I error and ANOVA overcomes this problem.
3. ANOVA plays an important role in multiple linear regression model diagnostics. The overall significance of the model is tested using ANOVA.
4. In a two-way ANOVA we check the impact of more than one factor simultaneously on several groups.

## MULTIPLE CHOICE QUESTIONS

1. For a one-way ANOVA, which of the following assumptions should be satisfied:
  - (a) The samples are drawn from a normal population.
  - (b) The response variable should be a continuous variable.
  - (c) The standard deviation of different groups should be equal.
  - (d) All of above

2. For an experiment with a single factor with  $k$  levels with  $n$  observations, the degrees of freedom for sum of squares of variation within the group is  
 (a)  $n - 1$       (b)  $k - 1$       (c)  $n - k$       (d)  $n - k + 1$
3. For a one-way ANOVA, the hypothesis test is a  
 (a) Right-tailed test      (b) Left-tailed test  
 (c) Two-tailed test      (d) Depends on null hypothesis
4. A data scientist is studying the impact of marital status (single, married, and divorced) on the annual income. The sample contains 140 singles, 110 married, and 40 divorced people. The values of  $SSB = 2425.6$  and  $SSW = 49,567.8$ . The value of  $F$ -statistic is  
 (a) 20.43      (b) 0.0489      (c) 7.02      (d) 0.1424
5. In a two-way ANOVA  
 (a) The number of factors is two      (b) The number of levels in a factor is two  
 (c) The number of factors is more than two      (d) The number of levels under each factor is more than two

### EXERCISES

1. If 10  $t$ -tests are conducted at  $\alpha=0.05$  and all are statically significant, calculate the value of Type I error that all tests are simultaneously significant.
2. Ms Sophia Smith, Senior Manager of Career Development Services (CDS) at the Institute of Science and Business (ISB), believes that the salary of graduating MBA students depends on their degree of specialization. To test her belief, Ms Smith collected discipline-wise graduating annual salary (in millions of rupees) from 2016 graduating students and the data is shown in Table 7.9. Conduct a one-way ANOVA at  $\alpha=0.05$  to check whether the annual salary depends on the degree discipline.

**TABLE 7.9** Annual salary of graduating students in millions of rupees for different degree disciplines

Engineering									
1.79	2.34	2.83	2.52	1.92	1.72	2.33	2.08	1.84	2.12
2.20	2.76	1.81	2.42	1.66	2.14	2.93	2.03	2.60	2.02
2.11	2.39	1.92	2.35	2.55	2.82	1.81	2.45	2.56	2.13
1.79	2.61	1.32	1.95	2.47	1.91	2.36	2.43	2.04	2.35
Science									
2.91	2.19	1.72	2.02	1.36	1.84	1.88	1.75	2.04	1.76
2.32	2.11	1.86	1.86	1.97	2.76	2.62	1.61	1.58	1.57
2.20	1.61	1.56	1.59	1.86	2.56	1.55	1.90	1.47	2.12
Commerce									
1.58	2.42	2.55	1.79	1.91	0.82	0.63	2.14	1.21	2.65
2.24	2.11	1.83	1.68	2.06	0.51	2.92	2.53	1.27	2.70

3. An original equipment manufacturer of a washing machine is interested in finding the impact of three different technologies on the reliability of the washing machine. Data on time between failures (in number of days) of the washing machine manufactured using different technologies is shown in Table 7.10. Conduct a one-way ANOVA at  $\alpha=0.01$  to check whether the mean times between failures are different for different technologies.

**TABLE 7.10** Time between failures of washing machine under different technologies

Technology 1									
340	324	353	326	319	358	287	327	366	270
271	343	327	357	304	359	195	292	307	250
292	393	328	298	294	353	392	293	252	315
327	299	298	324	363	337	336	295	339	290
451	370	331	413	371	322	313	329	274	407
Technology 2									
369	385	362	334	296	360	330	360	353	345
352	360	275	357	363	329	346	404	403	325
Technology 3									
352	419	375	403	437	418	375	410	358	305
432	418	367	400	360	349	375	395	405	382
400	327	320	389	427	391	363	380	419	376

4. In continuation of Question 2, Ms Smith was told by Mr Dicki Bird, Chairman of the Career Development Services that one should also look at the work experience in addition to the degree discipline. Ms Smith grouped the students with less than 2 years of experience and more than 2 years of experience and collected a new set of data which is shown in Table 7.11. Conduct a two-way ANOVA at  $\alpha=0.05$  to check whether the factors – degree discipline and years of work experience – have an impact on the graduating salary.

**TABLE 7.11** Data related to salary, degree discipline and year of experience

	Engineering									
	1.57	1.26	1.53	1.45	1.26	1.52	1.36	1.54	1.97	1.95
Less than 2 years	1.6	1.64	0.76	1.38	2.16	0.84	1.68	1.66	1.77	1.02
	2.47	1.75	1.45	1.86	1.77	0.9	1.39	2.08	1.8	1.7
	2.09	1.77	2.07	2.15	1.18	2.18	1.89	1.63	2.14	2.61
More than 2 years	2.05	1.41	1.28	1.1	2.12	2.06	1.73	2.46	1.61	2.07
	2.15	1.8	2.53	2.09	2.65	2.51	1.57	1.63	1.99	2.07
	Science									
Less than 2 years	1.18	1.47	1.72	1.57	1.62	0.76	1.85	1.18	1.97	1.77
	1.43	1.44	0.98	1.16	1.75	1.09	1.31	1.3	1.55	1.29
More than 2 years	2.15	1.43	1.77	1.81	1.59	1.64	0.85	2.8	0.94	1.64
	1.88	2.11	1.43	1.69	2.1	1.69	1.81	2.18	2.04	1.59
	Commerce									
Less than 2 years	2.23	1.99	2.78	1.91	2.72	2.13	2.18	1.18	2.3	1.79
	2.09	2.03	2.18	2.27	1.09	2.25	1.6	2.1	2.21	1.37
More than 2 years	1.3	2.03	2.24	2.18	2.44	2.84	1.5	2.13	2.72	1.75
	1.58	2.12	2.46	2.43	1.96	1.55	1.95	1.87	2.72	1.82

**REFERENCES**

1. Fisher R A (1934), “*Statistical Methods for Research Workers*”, Oliver and Boyd, London.
2. Kirk R E (1995), “*Experimental Designs: Procedures for the Behavioural Sciences*”, 3<sup>rd</sup> Edition, Brooks Cole, New York.
3. Kutner M H, Nachtsheim N J, Nester J, and Li W (2013), “*Applied Linear Statistical Models*”, 5<sup>th</sup> Edition, McGraw Hill.
4. Lunney G H (1969), “Individual Size for Multiple *t*-Tests”, *American Educational Research Journal*, **6**(4), 701–703.
5. Siegel A F (1990), “Multiple *t*-Tests: Some Practical Considerations”, *TESOL Quarterly*, **24**(4), 773–775.



# 8

# Correlation Analysis

“Success is not the key to Happiness; Happiness is the key to success. If you love what you are doing you will be successful”.

— Albert Schweitzer

## LEARNING OBJECTIVES

- LO 8-1** Understand the concept of correlation and its role in analytics.
- LO 8-2** Learn to calculate correlation between two continuous variables.
- LO 8-3** Understand the difference between correlation and causation.
- LO 8-4** Understand correlation between a continuous variable and a discrete variable.
- LO 8-5** Learn to calculate correlation between two discrete variables.

## CORRELATION

Correlation is a statistical measure of an association relationship between two random variables.

IMPORTANT

*Correlation is not necessarily a causal relationship. Correlation is important in analytics since it helps to identify variables that may be used in the model building and also useful for identifying issues such as multi-collinearity that can destabilize regression-based models.*

## 8.1 | INTRODUCTION TO CORRELATION

One of the challenging tasks in analytics, especially in predictive analytics, is identifying the variables or features that may be associated to the response variable or the outcome variable that is of interest to the data scientists. Organizations collect data on several variables, sometimes the number of variables can run into thousands (including derived variables such as ratios and interactions). For example, mobile service providers collect data on variables such as call duration, number of calls, numbers to which the calls are made, number of calls received, the device that was used to make the call, location (and mobile tower that the phone was attached to), time between calls, last recharge (in case of pre-paid mobile services), recharge amount, service plan (in case of post-paid connection), number of messages sent, number of messages received, apps downloaded, time spent on surfing internet, and so on. The number of variables collected and new variables generated may exceed several thousands.

Few of these variables are regulatory requirements as the part of the internal security and the mobile phone service providers are expected to collect and store. The idea behind collecting all these variables is to find answer to questions such as

1. Which customer is likely to churn?
2. How to increase the revenue generated from a customer?
3. What is the customer lifetime value?
4. What is the best service plan for a customer?
5. What recommendations can be made to a customer?

Finding answer to the aforementioned questions involves building predictive/prescriptive analytics models. Model building involves identifying the variables among thousands of variables (in analytics terminology this is called variable selection or feature selection) to build the model. Taking all the variables simultaneously to create a model can result in problems such as multi-collinearity, which can destabilize the model and is also time consuming since most predictive analytics model development involves matrix operations such as matrix inverse calculation. So, the knowledge of how different variables are related to one another is important in building analytical models. Correlation is a measure of the strength and direction of relationship that exists between two random variables and is measured using correlation coefficient. In other words, correlation is a measure of association between two variables. Correlation can assist the data scientists to choose the variables for model building that is used for solving an analytics problem. We will be discussing different types of correlation coefficients in this chapter depending on the scale of measurement of the variables involved.



### IMPORTANT

*Correlation is only an association relationship and not a causal relationship. Thus the user should be aware of the fact that two variables may have high correlation coefficient value, although there may not be any direct dependence between these variables.*

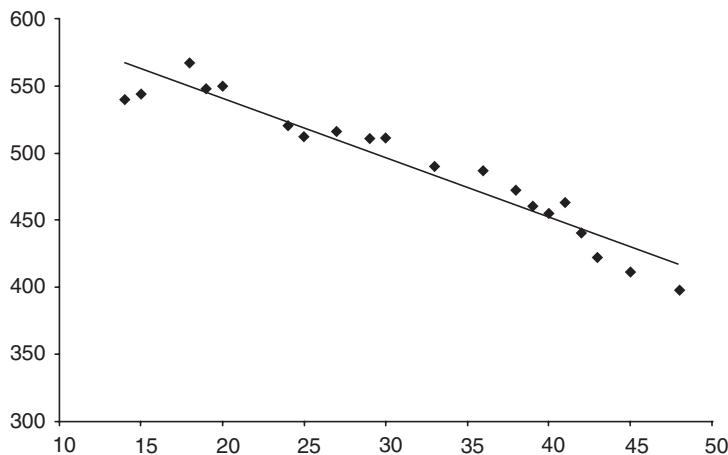
## 8.2 | PEARSON CORRELATION COEFFICIENT

Pearson product moment correlation (in short Pearson correlation) is used for measuring the strength and direction of the linear relationship between two continuous random variables  $X$  and  $Y$ . For example, consider two variables – the average call duration (variable  $Y$ ) and the age (variable  $X$ ). We may like to know whether the average call duration is related to the age of the caller, that is, whether change in age is related to change in average call duration. It is also possible that there may not be any relationship between age and average call duration. A simple approach for checking existence of association relationship is to draw a scatter plot. Scatter plot may reveal if there exists any relationship between two variables. In Table 8.1 we have age of customer and average call duration (measured in seconds) from a sample data; the corresponding scatter plot is shown in Figure 8.1.

In Figure 8.1, we can see that the average call duration ( $Y$ ) decreases as the age of the customer ( $X$ ) increases. We can measure the strength of the linear association relationship using a numerical measure

**TABLE 8.1** Data on age and average call duration (in seconds)

Age	14	15	18	19	20	24	25	27	29	30
Call Duration	540	544	567	548	550	520	512	516	511	511
Age	33	36	38	39	40	41	42	43	45	48
Call Duration	490	487	472	460	455	463	440	422	411	397

**FIGURE 8.1** Association relationship between age and average call duration.

called correlation coefficient. In the next section, we will be discussing mathematical equations for calculating Pearson product moment correlation coefficient.

### 8.2.1 | Calculation of Pearson Product Moment Correlation Coefficient

Pearson product moment correlation is used when we are interested in finding linear relationship between two continuous random variables (that is, the variable should be either of ratio or interval scale). When we try to measure how change in a variable (say  $Y$ ) is related to changes in another variable (say  $X$ ), one of the issues that we need to consider is the measurement scale and unit of measurement of the two variables. In the example discussed in Table 8.1, the variable age is measured in years and the call duration is measured in seconds. The range of two variables can be different, thus we need to standardize the variables which can be used for measuring the correlation between two variables.

Let  $X_i$  be different values of the variable  $X$  and  $Y_i$  be different values of  $Y$ . Then the standardized values of  $X$  and  $Y$  are given by

$$Z_X = \left( \frac{X_i - \bar{X}}{\sigma_X} \right) \quad (8.1)$$

$$Z_Y = \left( \frac{Y_i - \bar{Y}}{\sigma_Y} \right) \quad (8.2)$$

where  $\bar{X}$  and  $\bar{Y}$  are mean values of random variables  $X$  and  $Y$ ;  $\sigma_X$  and  $\sigma_Y$  are the corresponding standard deviations. The Pearson's correlation coefficient is given by

$$r = \frac{\sum_{i=1}^n Z_X Z_Y}{n} = \frac{\sum_{i=1}^n (X_i - \bar{X}) \times (Y_i - \bar{Y})}{n \sigma_X \sigma_Y} \quad (8.3)$$

where  $n$  is the number of cases in the sample. The formula in Eq. (8.4) is also frequently used to account for the degrees of freedom and recommended when the standard deviation is calculated from sample. For large samples, the correlation coefficients calculated using Eqs. (8.3) and (8.4) will converge.

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X}) \times (Y_i - \bar{Y})}{(n-1) S_X S_Y} \quad (8.4)$$

where  $S_X$  and  $S_Y$  are the standard deviations of random variables  $X$  and  $Y$  calculated from the sample. We can note the following properties from Eq. (8.3):

1. Whenever the value of  $X_i$  is greater than mean and if the corresponding value of  $Y_i$  is also greater than mean, then the numerator in equation will be positive.
2. Whenever the value of  $X_i$  is lesser than mean and if the corresponding value of  $Y_i$  is also lesser than mean, then the numerator in equation will be positive.
3. Whenever the value of  $X_i$  is lesser than mean (or greater than mean) and the corresponding value of  $Y_i$  is greater than mean (or lesser than mean), then the numerator in equation will be negative.

It is possible that we may have combinations of three cases listed above in a data set. Thus the numerator in Eq. (8.3) is likely to be positive, negative, or zero. The value of Pearson's correlation coefficient lies between  $-1$  and  $+1$ . Equation (8.3) is mathematically equivalent to Eqs. (8.5), (8.6), and (8.7):

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \times \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (8.5)$$

$$r = \frac{n \sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{\sqrt{n \sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2} \times \sqrt{n \sum_{i=1}^n Y_i^2 - (\sum_{i=1}^n Y_i)^2}} \quad (8.6)$$

$$r = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \quad (8.7)$$

where  $\text{Cov}(X, Y)$  is the covariance between random variables  $X$  and  $Y$  and is given by

$$\text{Cov}(X, Y) = E((X_i - \bar{X})(Y_i - \bar{Y})) \quad (8.8)$$

### Properties of Pearson Correlation Coefficient

1. The value of correlation coefficient lies between  $-1$  and  $+1$ . High absolute value of  $r$ ,  $|r|$ , indicates strong relationship between the two variables.
2. Positive value of  $r$  indicates positive correlation (as value of  $X$  increases, the value of  $Y$  also increases) and negative value of  $r$  indicates negative correlation (as the value of  $X$  increases, the value of  $Y$  decreases).
3. The sign of correlation coefficient is same as the sign of covariance between the two random variables.
4. Assume that the value of Pearson correlation coefficient between  $X$  and  $Y$  is  $r$ . Let  $Z_1$  and  $Z_2$  be the linear combinations of  $X$  and  $Y$  ( $Z_1 = A + BX$  and  $Z_2 = C + DY$ ). Then the correlation coefficient between  $Z_1$  and  $Z_2$  will be  $r$  when the signs of  $B$  and  $D$  are same (both are positive or negative) and  $-r$  when the signs of  $B$  and  $D$  are opposite.
5. Mathematically, square of correlation coefficient is equal to the co-efficient of determination ( $R^2$ ) of the linear regression model, that is  $r^2 = R^2$ .
6. Pearson correlation coefficient value may be zero even when there is a strong non-linear relationship between variables  $X$  and  $Y$  (Reed, 1917). Thus, low correlation coefficient value cannot be taken as an evidence of no relationship.



**IMPORTANT**

*Pearson correlation coefficient is a measure of linear relationship. Pearson correlation may not capture existence of non-linear relationship.*

### EXAMPLE 8.1

The average share prices of two companies over the past 12 months are shown in Table 8.2. Calculate the Pearson correlation coefficient.

**TABLE 8.2** Share prices (monthly average) of two companies over last 12 months

$X$	$Y$
274.58	219.50
287.96	242.92
290.35	245.90

**TABLE 8.2** Share prices (monthly average) of two companies over last 12 months—Continued

X	Y
320.07	256.80
317.40	240.60
319.53	245.23
301.52	232.09
271.75	222.65
323.65	231.74
259.80	214.43
263.02	201.86
286.03	204.23

The average values are  $\bar{X} = 292.9717$  and  $\bar{Y} = 229.8292$ .

The following equation is used for calculating the correlation coefficient:

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \times \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

The values are shown in Table 8.3.

**TABLE 8.3** Calculation of correlation coefficient

$X_i$	$Y_i$	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
274.58	219.50	-18.39	-10.33	189.97	338.25	106.6917
287.96	242.92	-5.01	13.09	-65.61	25.12	171.3699
290.35	245.90	-2.62	16.07	-42.13	6.87	258.2717
320.07	256.80	27.10	26.97	730.86	734.32	727.4259
317.40	240.60	24.43	10.77	263.11	596.74	116.0109
319.53	245.23	26.56	15.40	409.02	705.35	237.1857
301.52	232.09	8.55	2.26	19.33	73.07	5.111367
271.75	222.65	-21.22	-7.18	152.35	450.36	51.54043
323.65	231.74	30.68	1.91	58.62	941.16	3.651284
259.80	214.43	-33.17	-15.40	510.82	1100.36	237.1343
263.02	201.86	-29.95	-27.97	837.72	897.10	782.2743
286.03	204.23	-6.94	-25.60	177.70	48.19	655.3173
Sum				3241.77	5916.89	3351.98

From Table 8.3, we have

$$\sum_{i=1}^{12} (X_i - \bar{X})(Y_i - \bar{Y}) = 3241.77$$

$$\sum_{i=1}^{12} (X_i - \bar{X})^2 = 5916.89$$

$$\sum_{i=1}^{12} (Y_i - \bar{Y})^2 = 3351.98$$

$$\text{Correlation coefficient } r = \frac{3241.77}{\sqrt{5916.89} \times \sqrt{3351.98}} = 0.7279$$

In Microsoft Excel, CORREL(array 1, array 2) will give the Pearson product moment correlation value.

### 8.2.2 | Spurious Correlation

One of the major problem with correlation is the possibility of spurious correlation between two random variables which in many cases is caused due to some other latent variable (hidden variable) that influences both variables for which the correlation is calculated. Following are few examples of spurious correlation between two random variables:

- Crime rate versus ice cream sale:** It has been reported that the sale of ice cream and crime rates are positively correlated (Levitt and Dubner, 2009). Obviously, ice cream is not driving the crime rate. In this case the hidden variable is the temperature (summer increasing the ice cream sale) and also increasing crime (people on vacation and locked houses becomes easy target).
- Doctors and deaths:** Number of doctors is positively correlated with number of deaths in villages, that is, as the number of doctors increases, the deaths also increase. We can be sure that doctors are not causing the deaths to increase (Young, 2001).
- Divorce rate in Maine and per capita consumption of margarine:** The divorce rate in Maine was highly correlated with per capita consumption of margarine (based on data between 1999 and 2009). The correlation coefficient was 0.9926 (Source: [tylervigen.com](http://www.tylervigen.com)<sup>1</sup>).

### 8.2.3 | Hypothesis Test for Correlation Coefficient

For any two sets of data the Pearson correlation coefficient in Eq. (8.7) is most likely to give a value other than zero. Many thumb rules exist to group the correlation value as no correlation, low correlation, medium correlation, and high correlation (Monroe and Stuit, 1933). For example, correlation coefficient value of less than 0.2 is considered as negligible correlation and above 0.7 as high correlation (Monroe

<sup>1</sup> <http://www.tylervigen.com/spurious-correlations>

and Stuit, 1933). We would like to know what should be the minimum value of Pearson correlation coefficient before we can consider it as statistically significant, that is, whether there is a statistically significant correlation between two random variables. Let  $\rho$  be the population correlation coefficient. The null and alternative hypotheses are given by

$$\begin{aligned} H_0: \rho &= 0 \text{ (there is no correlation between two random variables)} \\ H_A: \rho &\neq 0 \text{ (there is a correlation between two random variables)} \end{aligned}$$

The sampling distribution of correlation coefficient  $r$  follows an approximate  $t$ -distribution with  $(n - 2)$  degrees of freedom ( $df$ ) where  $n$  is the number of cases in the sample for calculating the correlation coefficient. Two degrees of freedom are lost since we estimate two mean values from the data. The mean of the sampling distribution is  $\rho$  and the corresponding standard deviation is (Ezekiel, 1941)

$$\sqrt{\frac{1-r^2}{n-2}} \quad (8.9)$$

The  $t$ -statistic for null hypothesis is given by

$$t_{\alpha/2,n-2} = \frac{r-\rho}{\sqrt{\frac{1-r^2}{n-2}}} \quad (8.10)$$

When the null hypothesis is  $\rho = 0$ , the test statistic in Eq. (8.10) becomes

$$t_{\alpha/2,n-2} = r \sqrt{\frac{n-2}{1-r^2}} \quad (8.11)$$

### EXAMPLE 8.2

For Example 8.1, conduct the following two hypothesis tests at  $\alpha = 0.05$ :

- (a) The correlation between share prices of two companies is zero.
- (b) The correlation between share prices of two companies is at least 0.5.

**Solution:**

- (a) The null and alternative hypotheses are:

$$\begin{aligned} H_0: \rho &= 0 \\ H_A: \rho &\neq 0 \end{aligned}$$

The corresponding  $t$ -statistic is

$$t = r \sqrt{\frac{n-2}{1-r^2}} = 0.7279 \sqrt{\frac{12-2}{1-0.7279^2}} = 3.3569$$

Note that this is a two-tailed test and the critical  $t$ -value at  $\alpha = 0.05$  and  $df = 10$  is 2.2281 [which can be obtained using the Excel function TINV(0.05, 10)]. Since the calculated  $t$ -statistic is higher than the critical  $t$ -value, we reject the null hypothesis and conclude that there is a significant correlation between share prices of two companies. The corresponding  $p$ -value is 0.0072 [in Excel T.DIST.2T(3.3569, 10) = 0.0072].

(b) The null and alternative hypotheses are given by

$$\begin{aligned} H_0: \rho &\leq 0.5 \\ H_A: \rho &> 0.5 \end{aligned}$$

The corresponding  $t$ -statistic is

$$t = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}} = \frac{0.7279 - 0.5}{\sqrt{\frac{0.2168}{10 - 2}}} = 1.05$$

This is a right-tailed test and the corresponding  $t$ -critical value is 1.8124 [corresponding Excel function is TINV(0.1, 10)]. The calculated  $t$ -value is less than the critical value of  $t$ , and thus we retain the null hypothesis and conclude that the correlation between share prices of two companies is less than or equal to 0.5. The corresponding  $P$ -value is 0.1592 [T.DIST.RT(1.05, 10) = 0.1592].

### 8.3 | SPEARMAN RANK CORRELATION

Pearson correlation is appropriate when the random variables involved are both from either ratio scale or interval scale. When both random variables are of ordinal scale, we use Spearman rank correlation (also known as Spearman's rho denoted by  $\rho_s$ ). One of the problems with ordinal scale is that the data may be ranked based on the values, but the difference in ranks or ratio of ranks within each random variable will not be meaningful. The Spearman rank correlation,  $r_s$ , estimated from a sample with distinctive integer ranks is given by (Yule and Kendall 1937, Woodbury, 1940)

$$r_s = 1 - \frac{6 \sum_{i=1}^n D_i^2}{n(n^2 - 1)} \quad (8.12)$$

where  $D_i$  = difference in the rank of case  $i$  under variables  $X$  and  $Y$  (that is  $X_i - Y_i$ ). The sampling distribution of Spearman correlation  $r_s$  also follows an approximate  $t$ -distribution with mean  $\rho_s$  and standard

deviation  $\sqrt{\frac{1 - r_s^2}{n - 2}}$  with  $n - 2$  degrees of freedom.

**EXAMPLE 8.3**

Ranking of 12 countries under corruption and Gini Index (wealth discrimination) are shown in Table 8.4. Calculate the Spearman correlation and test the hypothesis that the correlation is at least 0.2 at  $\alpha = 0.02$ .

**TABLE 8.4** Ranking of countries under corruption and Gini Index

Country	1	2	3	4	5	6	7	8	9	10	11	12
Corruption	1	4	12	2	5	8	11	7	10	3	6	9
Gini Index	2	3	9	5	4	6	10	7	8	1	11	12

**Solution:**

The Spearman rank correlation calculations are shown in Table 8.5.

**TABLE 8.5** Spearman correlation calculation

Country	Corruption Rank ( $X_i$ )	Gini Index ( $Y_i$ )	$D = X_i - Y_i$	$D^2$
1	1	2	-1	1
2	4	3	1	1
3	12	9	3	9
4	2	5	-3	9
5	5	4	1	1
6	8	6	2	4
7	11	10	1	1
8	7	7	0	0
9	10	8	2	4
10	3	1	2	4
11	6	11	-5	25
12	9	12	-3	9
$\sum_{i=1}^{12} D_i^2$				68

The Spearman rank correlation is given by

$$r_s = 1 - \frac{6 \sum_{i=1}^n D_i^2}{n(n^2 - 1)} = 1 - \frac{6 \times 68}{12(12^2 - 1)} = 0.7622$$

The null and alternative hypotheses are

$$H_0: \rho_s \leq 0.2$$

$$H_A: \rho_s > 0.2$$

The corresponding  $t$ -statistic is

$$t = \frac{r_s - \rho_s}{\sqrt{\frac{1 - r_s^2}{n - 2}}} = \frac{0.7622 - 0.2}{\sqrt{\frac{0.2046}{10 - 2}}} = 2.74$$

The one-tailed  $t$ -critical value for  $\alpha = 0.02$  and  $df = 10$  is 2.35. Since the calculated  $t$ -statistic value is more than the  $t$ -critical value, we reject the null hypothesis and conclude that Spearman rank correlation between two countries is at least 0.2.

## 8.4 | POINT BI-SERIAL CORRELATION

Point bi-serial correlation is used when we are interested in finding correlation between a continuous random variable and a dichotomous (binary) random variable. Assume that the random variable  $X$  is a continuous random variable and  $Y$  is a dichotomous random variable. Then the following steps are used for calculating the correlation between these two variables:

1. Group the data into two sets based on the value of the dichotomous variable  $Y$ . That is, assume that the value of  $Y$  is either 0 or 1. Then we group the data into two subsets such that in one group the value of  $Y$  is 0 and in another group the value of  $Y$  is 1.
2. Calculate the mean values of two groups: Let  $\bar{X}_0$  and  $\bar{X}_1$  be the mean values of groups with  $Y = 0$  and  $Y = 1$ , respectively.
3. Let  $n_0$  and  $n_1$  be the number of cases in a group with  $Y = 0$  and  $Y = 1$ , respectively, and  $S_x$  be the standard deviation of the random variable  $X$ .

The point bi-serial correlation is given by (Pearson, 1909 and Soper, 1914)

$$r_b = \frac{\bar{X}_1 - \bar{X}_0}{S_x} \sqrt{\frac{n_0 n_1}{n(n-1)}} \quad (8.13)$$

where  $n$  is the total number of cases in the sample and  $S_x$  is the standard deviation of  $X$  estimated from sample and is given by

$$S_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$

### EXAMPLE 8.4

Ms Sandra Ruth, data scientist at Airmobile, is interested in finding the correlation between the average call duration and gender. Table 8.6 provides the average call duration (measured in seconds) and gender of 30 customers of Airmobile. In Table 8.6, male is coded as 0 and Female is coded as 1. Calculate the point bi-serial correlation.

**TABLE 8.6** Data on average call duration and gender

Gender	1	1	0	0	0	1	0	1	1	0
Call Duration	448	335	210	382	407	231	359	287	288	347
Gender	1	1	1	1	1	0	0	1	0	0
Call Duration	408	382	303	201	447	439	383	277	279	213
Gender	1	1	0	1	1	0	1	0	1	0
Call Duration	383	355	362	401	331	421	367	437	326	351

**Solution:**

From the data, we can calculate the following values:

$$\bar{X} = 345.33, \bar{X}_0 = 353.07, \bar{X}_1 = 339.4118, S_x = 71.7189, n_0 = 13, n_1 = 17$$

Bi-serial correlation is given by

$$r_b = \frac{\bar{X}_1 - \bar{X}_0}{S_x} \sqrt{\frac{n_0 n_1}{n(n-1)}} = \frac{339.4118 - 353.07}{71.7189} \sqrt{\frac{13 \times 17}{30(29)}} = -0.0960$$

There is very low negative correlation between gender and call duration.

## 8.5 | THE PHI-COEFFICIENT

Karl Pearson recommended the use the Phi-coefficient when both variables are binary for calculating the association relationship (Cramer, 1946). Let  $X$  and  $Y$  be two random variables both taking binary values (that is,  $X$  takes values 0 or 1 and similarly  $Y$  also takes values either 0 or 1). One can create a contingency table as shown in Table 8.7.

**TABLE 8.7** Contingency table

	$\gamma=0$	$\gamma=1$	Total
$X=0$	$N_{00}$	$N_{01}$	$N_{x0} = N_{00} + N_{01}$
$X=1$	$N_{10}$	$N_{11}$	$N_{x1} = N_{10} + N_{11}$
Total	$N_{y0} = N_{00} + N_{10}$	$N_{y1} = N_{01} + N_{11}$	

In the contingency table (Table 8.7):

$N_{00}$  = Number of cases in the sample such that  $X = 0$  and  $Y = 0$

$N_{01}$  = Number of cases in the sample such that  $X = 0$  and  $Y = 1$

$N_{10}$  = Number of cases in the sample such that  $X = 1$  and  $Y = 0$

$N_{11}$  = Number of cases in the sample such that  $X = 1$  and  $Y = 1$

$N_{x0}$  = Number of cases in the sample such that  $X = 0$

$N_{x1}$  = Number of cases in the sample such that  $X = 1$

$N_{y0}$  = Number of cases in the sample such that  $Y = 0$

$N_{y1}$  = Number of cases in the sample such that  $Y = 1$

The Phi-coefficient is given by

$$\phi = \frac{N_{11}N_{00} - N_{10}N_{01}}{\sqrt{N_{x0}N_{x1}N_{y0}N_{y1}}} \quad (8.14)$$

**EXAMPLE 8.5**

Joy Finance (JF) is a company that provides gold loans (in which gold is used as guarantee against the loan). Mr Georgekutty, Managing Director of JF, collected data to understand the relationship between loan default status (variable  $Y$ ) and the marital status of the customer (variable  $X$ ). Data is collected on past 40 loans and is shown in Table 8.8. Calculate the Phi-coefficient. In Table 8.8,  $Y = 0$  implies non-defaulter,  $Y = 1$  is defaulter,  $X = 0$  is single, and  $X = 1$  is married.

**TABLE 8.8** Marital status (0 = Single, 1 = Married) versus loan status (0 = No default, 1 = Default)

$X$	1	0	1	0	0	0	0	0	1	0
$Y$	0	1	0	1	0	0	0	1	1	1
$X$	0	1	1	0	0	1	0	0	0	1
$Y$	0	1	1	1	0	0	1	1	0	0
$X$	1	0	0	0	1	1	1	0	0	1
$Y$	0	0	0	1	0	0	0	0	0	0
$X$	1	0	0	0	1	0	1	0	1	1
$Y$	1	0	0	1	1	0	1	1	0	1

**Solution:**

The contingency table for the data shown in Table 8.8 is given in Table 8.9.

**TABLE 8.9** Contingency table

		$Y$		Total
		0	1	
$X$	0	13	10	23
	1	10	7	17
Total		23	17	40

From Table 8.9, we have

$$N_{00} = 13, N_{01} = 10, N_{10} = 10, N_{11} = 7, N_{x0} = 23, N_{x1} = 17, N_{y0} = 23, \text{ and } N_{y1} = 17$$

The Phi-coefficient is given by

$$\phi = \frac{N_{11}N_{00} - N_{10}N_{01}}{\sqrt{N_{x0}N_{x1}N_{y0}N_{y1}}} = \frac{7 \times 13 - 10 \times 10}{\sqrt{23 \times 17 \times 23 \times 17}} = -0.0230$$

Since the Phi-coefficient is very small, we can conclude that there is not much correlation between the marital status and loan default.

**SUMMARY**

1. Correlation is a measure of strength and direction of linear relationship between two random variables. It can be used only when the relationship is linear.
2. Correlation captures only association relation and not a causal relation.
3. Pearson product moment correlation is used when two random variables are continuous. In the case of two ordinal variables, the appropriate correlation is Spearman rank correlation. Bi-serial correlation is used when correlation is calculated between one continuous and one binary random variable. Phi-coefficient is used for calculating correlation between two binary random variables. Bi-serial correlation and Phi-coefficient have the same interpretation as Pearson correlation coefficient.
4. Correlation is an important measure and can be used for feature selection while building regression models.
5. One of the drawbacks of correlation is the spurious correlations. It is possible that two variables with no explainable relationship may have high correlation coefficient.

**MULTIPLE CHOICE QUESTIONS**

1. Pearson product moment correlation is used when
  - (a) Both random variables are from ordinal scale
  - (b) Both random variables are from either interval scale or ratio scale
  - (c) Both random variables are from nominal scale
  - (d) Both random variables are from ratio scale
2. If the Pearson correlation coefficient between two variables is positive then
  - (a) The covariance between the variables is negative
  - (b) The covariance between the variables is positive
  - (c) The standardized values of both random variables are always positive
  - (d) None of the above
3. Which of the following statements are correct?
  - (a) Correlation establishes dependency relationship
  - (b) Correlation establishes causal relationship
  - (c) Correlation establishes association relationship
  - (d) All of above
4. Pearson correlation between two variables  $X$  and  $Y$  is 0.85. Pearson correlation between  $2 + 3X$  and  $4 - 5Y$  is
 

(a) Less than 0.85	(b) More than 0.85	(c) 0.85	(d) -0.85
--------------------	--------------------	----------	-----------
5. Covariance between two random variables is 0.5; correlation between these two random variables will be
 

(a) At least 0.5	(b) At most 0.5	(c) 0.25	(d) 0.75
------------------	-----------------	----------	----------

**EXERCISES**

1. Professor Bell at Bellandur University, Bangalore believes that the cumulative grade point average (CGPA) of the students is negatively correlated with usage (measured in average minutes per day) of smart phones. Table 8.10 shows the CGPA and smart phone usage in minutes per day of 40 students.
  - (a) Calculate the Pearson correlation coefficient between CGPA and mobile phone usage of students.
  - (b) Conduct a hypothesis test at  $\alpha = 0.01$  to check whether CGPA and mobile phone usage are negatively correlated.
  - (c) Professor Bell believes that the correlation is less than -0.4. Conduct a hypothesis test at  $\alpha = 0.1$  to check whether the claim is correct.

**TABLE 8.10** Data of CGPA and mobile phone usage (average minutes per day)

CGPA	2.65	2.25	1.86	1.47	2.10	1.94	2.71	1.83	2.65	2.04
Phone usage	75	89	65	136	95	103	74	109	75	98
CGPA	2.54	2.16	2.28	2.47	2.18	2.57	1.97	2.87	2.10	3.28
Phone usage	60	93	88	81	92	78	102	70	95	89
CGPA	2.78	2.44	1.87	2.50	2.24	2.01	2.17	2.20	2.05	1.63
Phone usage	72	82	107	80	89	100	92	91	98	123
CGPA	2.28	2.63	2.86	2.24	2.44	2.69	2.22	3.07	1.77	3.03
Phone usage	88	76	70	89	82	74	90	65	113	66

2. Mr Chellappa is the founder of Oho Productions that produces movies in different languages of India. Mr Chellappa believes that the length of the movie (measured in minutes) is not related to its box-office collection. Table 8.11 shows length of the movie (in minutes) and the box-office collection (in millions of rupees). Use an appropriate hypothesis test to check whether there is a correlation between length of the movie and the box-office collection at a significance level of 0.05.

**TABLE 8.11** Data on length of the movie and the box-office collection

Length of the movie	121	79	170	160	77	147	115	76	110	141
Box-office collection	1078	415	441	1192	258	1185	139	427	309	411
Length of the movie	100	82	82	114	110	163	92	172	142	136
Box-office collection	506	441	595	1728	1507	518	1463	1356	1014	422
Length of the movie	143	108	154	140	177	97	106	163	142	115
Box-office collection	508	1262	1783	1281	1253	1178	1103	454	301	296

3. Table 8.12 provides ranking of Indian states based on corruption and Table 8.13 provides ranking based on literacy rate. Calculate the Spearman rank correlation between the corruption rank and literacy rank.

**TABLE 8.12** Rank based on corruption (1 implies high corruption)

State	Bihar	Jammu and Kashmir	Madhya Pradesh	Uttar Pradesh	Karnataka	Rajasthan	Tamil Nadu	Chhattisgarh
Rank	1	2	3	4	5	6	7	8
State	Delhi	Gujarat	Jharkhand	Kerala	Orissa	Andhra Pradesh	Haryana	Himachal Pradesh
Rank	9	10	11	12	13	14	15	16

**TABLE 8.13** Rank based on literacy rate (1 implies high literacy)

State	Bihar	Jammu and Kashmir	Madhya Pradesh	Uttar Pradesh	Karnataka	Rajasthan	Tamil Nadu	Chhattisgarh
Rank	16	12	10	11	7	15	4	9
State	Delhi	Gujarat	Jharkhand	Kerala	Orissa	Andhra Pradesh	Haryana	Himachal Pradesh
Rank	2	5	13	1	8	14	6	3

Conduct a hypothesis test to check whether corruption and literacy rate are negatively correlated at  $\alpha = 0.05$ .

4. Harrison Seth, Dean of a Business School, believes that the outgoing salary of their MBA students may be correlated with their undergraduate specialization. Harrison believes that the students with engineering specialization at the undergraduate degree received more salary compared to other degrees. Table 8.14 shows the outgoing salary (in millions of rupees) of MBA graduates and their discipline in undergraduate (1 = engineering and 0 = non-engineering). Calculate the correlation between salary and engineering discipline,

**TABLE 8.14** Salary (in millions of rupees) and undergraduate degree

(1 = engineering and 0 = non-engineering)									
Degree	0	1	0	1	0	0	1	0	0
Salary	3.3	2.22	1.82	2.55	1.84	2.53	2.87	2.39	2.32
Degree	1	1	0	1	0	0	1	1	0
Salary	2.22	2.31	2.05	2.04	1.7	2.28	2.56	3.13	2.26
Degree	0	0	0	0	1	0	0	0	1
Salary	2.03	1.45	1.62	0.92	2.31	2.37	1.59	2.56	3.13

5. Telepower is a telephone service provider which collects data on customer churn and the number of mobile handsets used by the customer. Table 8.15 shows the data in which  $Y$  denotes churn ( $Y = 1$  implies churn and  $Y = 0$  implies no churn) and variable  $X$  denotes the number of handsets used by the customer where  $X = 0$  implies the customer uses single handset and  $X = 1$  implies the customer uses more than one handset for making phone calls. Calculate the Phi-coefficient for the data shown in Table 8.15.

**TABLE 8.15** Number of handsets ( $X$ ) and customer churn ( $Y$ )

$X$	1	1	0	0	0	1	1	1	1
$Y$	1	1	1	1	0	0	1	0	1
$X$	0	1	1	1	1	0	0	1	1
$Y$	0	1	0	1	1	0	0	1	1
$X$	1	1	1	0	1	0	1	0	1
$Y$	0	1	1	0	1	0	0	1	1
$X$	1	1	1	1	0	1	1	0	1
$Y$	0	1	0	1	1	1	1	0	1
$X$	0	0	1	0	1	0	1	1	0
$Y$	0	0	1	1	1	0	1	1	1

**REFERENCES**

1. Anon (2009), "Spurious Correlations", Tylervigen.com available at <http://www.tylervigen.com/spurious-correlations>, accessed on 2 April 2017.
2. Cramer H (1946), "Mathematical Methods of Statistics", Princeton University Press, NJ.
3. Ezekiel M (1941), "Methods of Correlation Analysis", Wiley, New York.
4. Levitt S D and Dubner S J (2009), "Super Freakonomics", Penguin Press, London.
5. Monroe W S and Stuit D B (1933), "The Interpretation of Coefficient of Correlation", *The Journal of Experimental Education*, 1(3), 186–203.
6. Pearson K (1909), "On a new method of determining correlation between a measured character A and a Character B of which only the percentage of cases wherein B exceeds (or falls short of) a given intensity is recorded for each grade of A", *Biometrika*, 7(1), 96–105.
7. Reed W G (1917), "The Correlation Coefficient", *Publication of the American Statistical Association*, 15(118), 670–684.
8. Soper H E (1914), "On the Probable Error of the Bi-Serial Expression for the Correlation Coefficient", *Biometrika*, 10(2/3), 384–390.
9. Woodbury M A (1940), "Rank Correlation when there are Equal Variates", *The Annals of Mathematical Statistics*, 11(3), 358–362.
10. Young F W (2001), "An Explanation of the Persistent Doctor Mortality Association", *Journal of Epidemiology and Community Health*, 55, 80–84.
11. Yule G U and Kendall M G (1937), "Introduction to the Theory of Statistics", Charlie Griffin, London.



# Simple Linear Regression

9

“If you torture the data long enough, it will confess.”

— Ronald Coase

## LEARNING OBJECTIVES

- LO 9-1** Learn fundamentals of simple linear regression and its applications in predictive analytics.
- LO 9-2** Understand the difference between causal relationship and association relationship.
- LO 9-3** Understand various stages in regression model building, and the underlying assumptions of regression models.
- LO 9-4** Learn method of Ordinary-Least-Squares (OLS) for estimation of regression parameters.
- LO 9-5** Interpretation of regression parameters under different functional forms.
- LO 9-6** Learn to carry out regression model diagnostics and validate the model.
- LO 9-7** Application of simple linear regression model in predictive analytics problems.

## ESSENCE OF SIMPLE LINEAR REGRESSION

Simple linear regression is a statistical technique for finding the **existence of an association relationship** between a dependent variable (aka response variable or outcome variable) and an independent variable (aka explanatory variable or predictor variable). Simple linear regression implies that there is only one independent variable in the model. Regression models do not establish causal relationship between the dependent variable (say  $Y$ ) and the independent variable ( $X$ ). Regression, however, can be used to check whether there is an association relationship between the variable  $Y$  and the variable  $X$ . That is, using regression, we cannot say that the value of the dependent variable  $Y$  depends on the value of the independent variable  $X$  (or a change in the value of  $Y$  is caused due to a change in the value of  $X$ ). We can only establish that the change in value of  $Y$  is associated with the change in value of  $X$ . For example, in the cricket test series that was played in England in 2014, India lost to England 3 to 1. Immediately after the defeat, Indian media reported that the Board for Control for Cricket in India (BCCI) banned girlfriends and wives from travelling with the players during tours, thus blaming

## ESSENCE OF SIMPLE LINEAR REGRESSION

the company of girlfriends (and wives) for the poor performances of the players. BCCI assumed that the cause of poor performance of the players in the test series was due to the presence of girlfriends and not the players' inability to adapt to the English conditions. One has to use other statistical techniques such as **Rubin Causal Model** to establish causal relationship between two variables of interest.

**IMPORTANT**

*Regression establishes existence of an association relationship between two variables, and not a causal relationship. The use of the term 'dependent variable' does not imply that the changes in the values of that variable are dependent on the changes in the values of the independent variable(s).*

### 9.1 | INTRODUCTION TO SIMPLE LINEAR REGRESSION

Regression is one of the most important techniques in predictive analytics since many prediction problems are modelled using regression. It is one of the supervised learning algorithms, that is, a regression model requires the knowledge of both the dependent and the independent variables in the training data set. Organizations use several key performance indicators (KPIs) such as cost of goods sold, customer lifetime value, growth rate, market share, productivity, profit, return on investment (ROI), etc. to measure their performance. KPIs, in turn, may be influenced by several factors. For example, the market share of a product sold by a company may be associated with factors such as:

1. Price of the product
2. Promotion expenses
3. Competitors' price
4. Competitor's promotion expenses
5. New product introductions
6. Macro-economic variables such as gross domestic product (GDP), inflation, unemployment, and so on.

Organizations would like to understand the existence of relationships between key performance indicators and other factors which may be exploited to improve the performance and to add value to the organization.

**Simple Linear Regression** (SLR) is a statistical model in which there is only one independent variable and the functional relationship between the dependent variable and the regression coefficient is linear.

One of the functional forms of SLR is given as follows:

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (9.1)$$

For a data set with  $n$  observations  $(X_i, Y_i)$ , where  $i = 1, 2, \dots, n$ , the functional form (9.1) can be written as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (9.2)$$

where

$Y_i$  is the value of  $i^{\text{th}}$  observation of the dependent variable in the sample

$X_i$  is the value of  $i^{\text{th}}$  observation of the independent variable in the sample

$\varepsilon_i$  is the random error (also known as residuals)

$\beta_0$  and  $\beta_1$  are the regression parameters (or regression coefficients)

The regression relationship stated in Eq. (9.2) is a statistical relationship, and thus is not exact, unlike a mathematical relationship, and thus the error terms  $\varepsilon_i$ . The dependent variable  $Y_i$  is often known as **response variable** or **outcome variable**, and the independent variable  $X_i$  is also known as **predictor variable** or **explanatory variable**.

SLR attempts to explain the changes in the value of response variable  $Y$  using the knowledge of the values of explanatory variables  $X$ . Thus, the equation  $\beta_0 + \beta_1 X_i$  gives the predicted value of  $Y_i$  for a given value of  $X_i$ , whereas the term  $\varepsilon_i$  is the error in predicting the values of  $Y_i$ . In fact  $\beta_0 + \beta_1 X_i$  is the conditional expected value of  $Y_i$  for a given value of  $X_i$ .

It is important to note that the linearity condition in linear regression is defined with respect to the regression coefficients ( $\beta_0$  and  $\beta_1$ ), and not with respect to the explanatory variables in the model. For example, the functional forms in Eqs. (9.3) and (9.4) are linear regression models, although the relationships between  $Y$  and  $X$  are not linear:

$$Y_i = \beta_0 + \beta_1 X_i^2 + \varepsilon_i \quad (9.3)$$

$$Y_i = \beta_0 + \beta_1 \ln(X_i) + \varepsilon_i \quad (9.4)$$

On the other hand, the regression relationships in Eqs. (9.5) and (9.6) are examples of non-linear regression models:

$$Y_i = \beta_0 + \frac{1}{1 + \beta_1} X_i + \varepsilon_i \quad (9.5)$$

$$Y_i = \beta_0 + e^{\beta_1} X_i + \varepsilon_i \quad (9.6)$$

Equations (9.5) and (9.6) are called **non-linear regression** since the relationship between the dependent variable  $Y$  and the regression coefficient  $\beta_1$  is non-linear. Thus estimating the regression coefficients will involve solving a system of non-linear equations. Many non-linear regression relationships can be converted into linear regression model by applying suitable transformations.

**IMPORTANT**

*Linear regression means that the relationship between the response variable ( $Y$ ) and the regression coefficient ( $\beta$ ) is linear.*

**TABLE 9.1** Data extracted from Galton (1886)

Height in inches (Mid-Parents)	Deviation	Children
64.5	4	66.3
65.5	3	67.7
66.5	2	67.8
67.5	1	67.9
68.5	0	68.3
69.5	-1	68.5
70.5	-2	69
71.5	-3	70
72.5	-4	71

## 9.2 | HISTORY OF REGRESSION—FRANCIS GALTON'S REGRESSION MODEL

Sir Francis Galton, an English statistician, was the first person to coin the term '*Regression*', based on his research on hereditary properties of successive generations of sweet peas and humans. Galton (1886) published an article titled '*Regression towards Mediocrity in Hereditary Stature*', in the journal of the anthropological institute of Great Britain and Ireland. Herein he described his experiments with the size of produce of different sweet pea seed sizes as well as the height of parents and their children.

Galton wanted to understand if any relationship existed between height of the parents and their children. For this, he collected data of 905 adult children from 205 families. Galton then transmuted female heights to the male equivalent by multiplying the female heights by a factor of 1.08 to account for the difference in the male and female heights. He observed that when the mid-parent height (average height of parents) was higher than the mean value of that generation, then their children tended to be shorter. However, when the parents were shorter than the mean value, their children tend to be taller.

The summary data and the graph based on the data collected by Galton are shown in Table 9.1 and Figure 9.1, respectively. From Table 9.1, it is evident that shorter parents had taller children (64.5 and 66.3, respectively), whereas taller parents had shorter children (72.5 and 71, respectively). Sir Francis Galton coined the term '*regression*' since the heights of children of taller parents were regressing towards the mean value of the offspring population.

Francis Galton probably never realized that the term that he coined would become one of the most important tool in analytics and econometrics.

## 9.3 | SIMPLE LINEAR REGRESSION MODEL BUILDING

A simple linear regression model is developed to understand how the value of a KPI is associated with changes in the values of an independent variable.

Some examples are as follows:

1. A hospital may be interested in finding how the total treatment cost of a patient varies with the body weight of the patient.

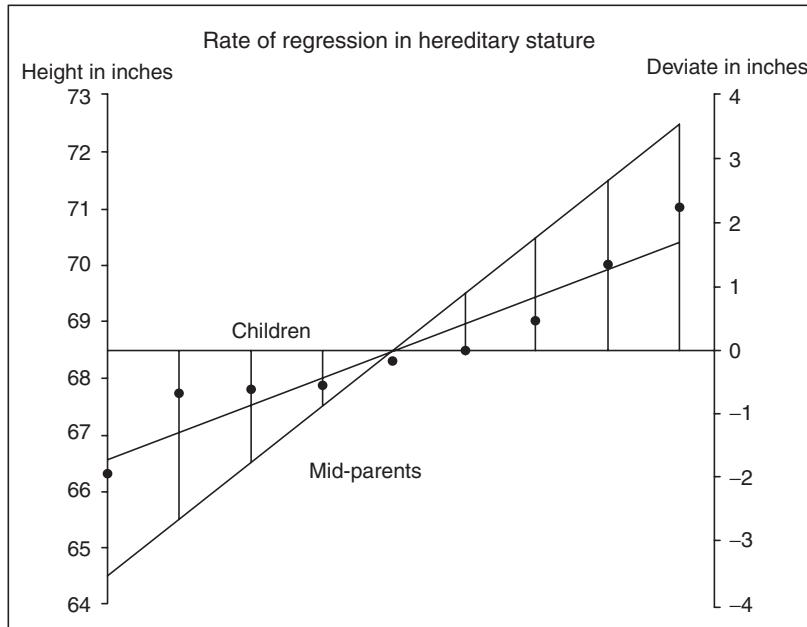
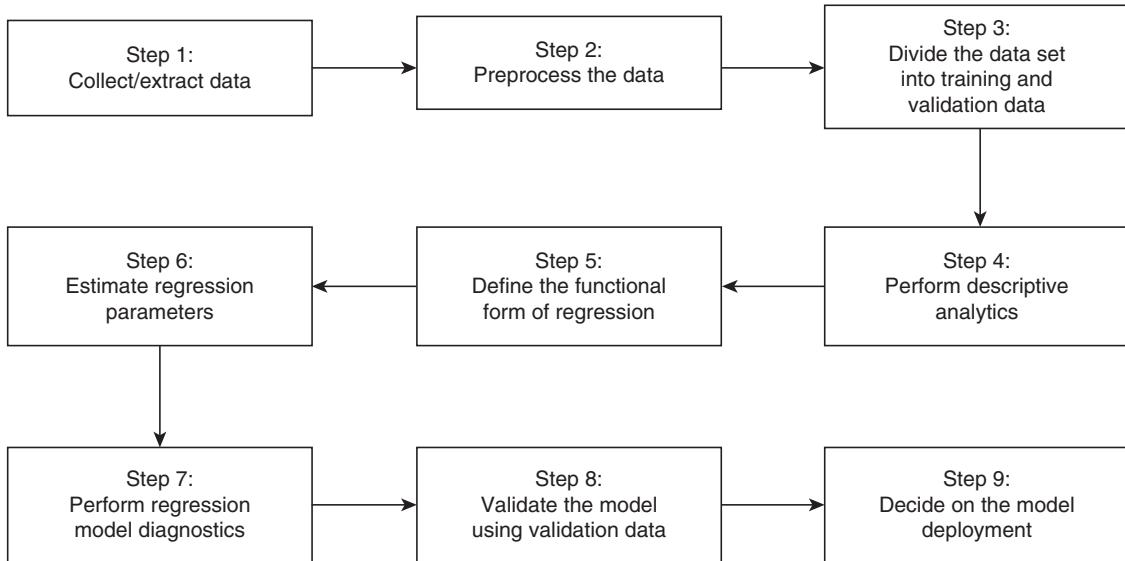


FIGURE 9.1 Francis Galton's regression graph (reproduced from Galton, 1886).

2. A business school may like to know the relationship between the salary offered to the graduating students by recruiting companies during their placements and their cumulative grade point average (CGPA).
3. Government (policy makers) may like to understand the impact of policies such as Mahatma Gandhi National Rural Employment Guarantee Act (NREGA) on the wages of labour force within different sectors of the industry.
4. E-commerce companies such as Amazon, Bigbasket and Flipkart would like to understand the number of customer visits to their portal and the revenue.
5. Restaurants would like to know the relationship between the customer waiting time after placing the order and the net promoters score (NPS).
6. Manufacturing companies would be interested in understanding the relationship between the service level agreements and the supply chain costs.
7. Retailers such as Walmart, Target, Reliance Retail, Hyper City, etc. would be interested in understanding the impact of price cut promotions on the revenue of their private labels (store brands or house brands).
8. Banks and other financial institutions would like to understand the impact of unemployment rate on percentage of non-performing assets (NPA).
9. Policy-makers would like to understand the impact of demonetization on gross domestic product (GDP).
10. Original equipment manufacturers (OEMs) would like to know the impact of duration of warranty on the profit.

The potential applications of regression models are thus immense.

The framework for regression model building is described in Figure 9.2. Regression model building within the business analytics context is usually triggered by the need to understand the relationships between a KPI that is important to the performance of the organization and other factors. Regression models can lead to previously unknown relationships, thereby leading to new hypothesis.



**FIGURE 9.2** Framework for SLR model development.

Activities performed during different steps of the regression model building are described below:

#### STEP 1 Collect/Extract Data

The first step in building a regression model is to collect and/or extract data from different sources for the identified problem (or KPI). Data collection, however, can be time-consuming and expensive.

1. Since ERP (enterprise resource planning) systems that are used for collecting and storing the data are not designed for analytics applications (which is probably changing in the recent past), data extraction from the ERP systems becomes a time-consuming process.
2. Another issue with data collection is that it can be expensive. For example, let us assume that a modeller believes that the amount spent by a customer in a retail store would be dependent on the customer's income. However, not many customers may like to reveal their income, thus making it difficult and expensive to collect the required data. In general, data that is seen as too personal may be difficult to collect and verify. Also regulations across many countries forbid collecting/sharing sensitive personal information.

**STEP 2** Pre-process the Data

Pre-processing the data is an important stage in the regression model building and is required for many reasons. For example, Francis Galton multiplied the height of every female by 1.08 to make it comparable to the heights of males (Hanley, 2004).

Before the model is built, it is also essential to ensure the quality of the data for issues such as reliability, completeness, usefulness, accuracy, missing data, and outliers. Reliability of data is a major issue when the data is collected through surveys and user-generated data. For example, according to Huffington Post, 53% men and 63% women lie about their profile (height, weight, job, etc.) in online dating sites (Hodge, 2012). Use of unreliable and incorrect data to build models can lead to incorrect hypothesis.

Missing data is another frequently observed problem and the data scientists have to come up with strategies for handling missing data. Data imputation techniques may be used to deal with missing data. Use of descriptive statistics and visualization (such as box plot and scatter plot) may be used to identify the existence of outliers and variability in the data set.

An important aspect of data pre-processing is that the modeller has to decide how he/she would use the data in the model building. Many new variables (such as ratio of variables or product of variables) can be derived and can be used in model building. **Categorical variables (qualitative variables or nominal scale variables)** in the data set cannot be used directly in the model since it can create model misspecification. For example, the data collector may have used different codes to enter the marital status of a person [such as married male (coded as 1), married female (coded as 2), divorced male (coded as 3), divorced female (coded as 4), etc.]. Such categorical data has to be pre-processed using **dummy variables** before it can be used in the regression model. Use of dummy variables to incorporate categorical variables in regression model is discussed in Chapter 10. Data collection and data pre-processing can take significant time in regression model building (in fact in many analytics projects).

**STEP 3** Divide the Data into Training and Validation Data Sets

In this stage the data is divided into two subsets (sometimes more than two subsets): **training data set** and **validation data set**. The proportion of training data set is usually between 70% and 80% and the remaining data is treated as the validation data. The subsets may be created using random/stratified sampling procedure. The actual sampling technique used for creating the training and validation sets may, however, depend on the problem context and structure of the data. The training data is used for developing the model and the validation data is used for model validation and selection.

For example, assume that three different models (say  $M_1$ ,  $M_2$ , and  $M_3$ ) are developed using the training data set. There is no guarantee that the best performing model in the training data will also perform well in the validation data. The best performing model in the training data may over-fit the data and thus may not perform well in the validation data. The primary objective of partitioning the data is to select the best model based on its performance in the validation data and avoid possible over-fitting during model development. In many cases, the original data set may be partitioned into three partitions: *training*, *validation* and *test data*.

**STEP 4** Perform Descriptive Analytics

It is always a good practice to perform descriptive analytics before moving to predictive analytics model building. Descriptive statistics will help us to understand the variability in the model and visualization of the data through, say, a box plot which will show if there are outliers in the data. Another visualization technique, the scatter plot, may also reveal if there is any obvious relationship between the two variables under consideration.

**STEP 5** Define the Functional Form of Relationship

For better predictive ability (model accuracy) it is important to specify the correct functional form between the dependent variable and the independent variable. Scatter plots may assist the modeller to define the right functional form.

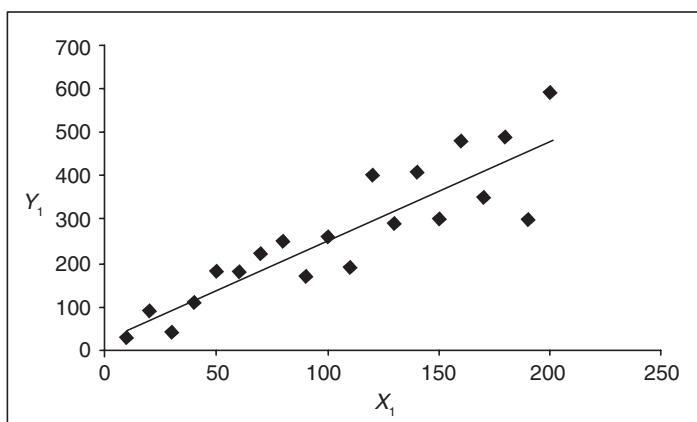


FIGURE 9.3 Linear relationship between  $X_1$  and  $Y_1$ .

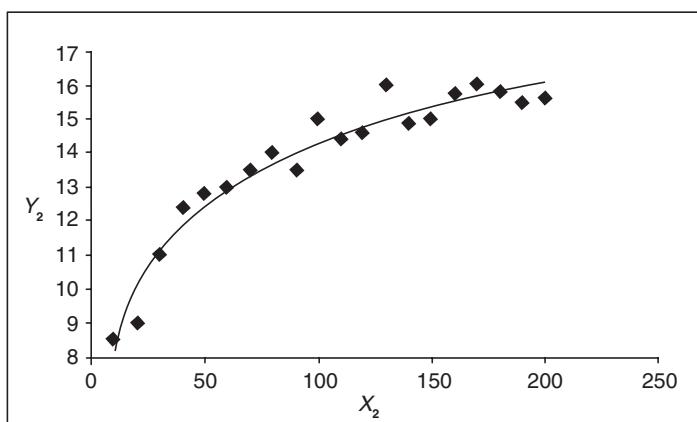


FIGURE 9.4 Log-linear relationship between  $X_2$  and  $Y_2$ .

For example, for the data shown in Figure 9.3, the functional form  $Y_1 = \alpha_0 + \alpha_1 X_1$  is appropriate, whereas for the data shown in Figure 9.4, the most appropriate functional form is  $\ln(Y_2) = \beta_0 + \beta_1 X_2$ . However, scatter plots can be inconclusive in many cases, especially when outliers are present in the data. Analysis of the residuals will be useful in case the modeller uses an incorrect functional form.

---

#### **STEP 6** Estimate the Regression Parameters

Once the functional form is specified, the next step is to estimate the regression parameters. The method of **Ordinary Least Squares (OLS)** is used to estimate the regression parameters.

OLS fits regression line through a set of data points such that the sum of the squared distances between the actual observations in the sample and the regression line is minimized [i.e.,  $\sum_i (Y_i - \hat{Y}_i)^2$  is minimized]. OLS provides the **Best Linear Unbiased Estimate (BLUE)**. That is,

$$E[\beta - \hat{\beta}] = 0$$

where  $\beta$  is the population parameter and  $\hat{\beta}$  is estimated parameter value from the sample (Waugh, 1961).

---

#### **STEP 7** Perform Regression Model Diagnostics

Regression is often misused since the modeller fails to perform necessary diagnostics tests before applying the model. Before it can be deployed it is necessary that the regression model created is validated for all model assumptions including the definition of the functional form. If the model assumptions are violated, then the modeller has to use some remedial measure; it is also possible that there is no association relationship between the variables. Remedial measures are necessary if any of the regression model assumptions are violated (the assumptions are discussed in Section 9.4). The remedial measures are discussed in Chapter 10.

---

#### **STEP 8** Validate the Model using the Validation Data Set

A major concern in analytics is over-fitting, that is, the model may perform very well in the training data set but may perform badly in validation data set. It is important to ensure that the model performance is consistent in the validation data set as was in the training data set. In fact, the model may be cross-validated using multiple training and test data sets.

---

#### **STEP 9** Decide on the Model Deployment

The final step in the regression model is to generate actionable items and business rules that can be used by the organization. For example, by predicting weather conditions using analytics model, farm advisory systems are created that can provide information and assistance to farmers. Note that farm

advisory systems are complex and are usually designed based on several complex regression models. Sample actionable items in retail could be decisions such as:

1. Deciding the price of a product.
2. Deciding the depth of price cut promotion (such as 10% or 20% discount).
3. Identifying the target customers and so on.

Note that the model will be deployed only when its performance is acceptable in the validation data set. In the next section, we will discuss the Ordinary Least Squares procedure used for estimation of the regression parameters.

---

## 9.4 | ESTIMATION OF PARAMETERS USING ORDINARY LEAST SQUARES

An important step in regression model building is the estimation of the regression parameters. Given a set of dependent variable values ( $Y_i$ ) and the corresponding independent variable values ( $X_i$ ), each subject to a random error ( $\varepsilon_i$ ), one has to find the best equation to represent the relationship between the dependent and independent variables.

The method of least squares gives the best equation under the assumptions stated below (Harter 1974, 1975):

1. The regression model is linear in regression parameters.
2. The explanatory variable,  $X$ , is assumed to be non-stochastic (i.e.,  $X$  is deterministic).
3. The conditional expected value of the residuals,  $E(\varepsilon_i | X_i)$ , is zero.
4. In case of time series data, residuals are uncorrelated, that is,  $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$  for all  $i \neq j$ .
5. The residuals,  $\varepsilon_i$ , follow a normal distribution.
6. The variance of the residuals,  $\text{Var}(\varepsilon_i | X_i)$ , is constant for all values of  $X_i$ . When the variance of the residuals is constant for different values of  $X_i$ , it is called **homoscedasticity**. A non-constant variance of residuals is called **heteroscedasticity**.



*Assumptions 3–6 are not necessary for the least square estimation. They are needed for the hypothesis tests that will be used for the validation of the regression model.*



*If the errors follow normal distribution, then  $Y$  will also follow a normal distribution, since  $\beta_0$  and  $\beta_1$  are parameters and  $X$  is deterministic.*

Assume that  $\{(Y_1, X_1), (Y_2, X_2), \dots, (Y_n, X_n)\}$  are data set with  $n$  observations. Using assumption (3), we can show that

$$E(Y_i | X_i) = E(\beta_0 + \beta_1 X_i + \varepsilon_i) = E(\beta_0) + E(\beta_1 X_i) + E(\varepsilon_i | X_i) = \beta_0 + \beta_1 X_i \quad (9.7)$$

Equation (9.7) follows from the fact that the  $\beta_0$  and  $\beta_1$  are parameters,  $X_i$  values are non-stochastic and  $E(\varepsilon_i|X_i) = 0$ . Thus,  $\beta_0 + \beta_1 X_i$  is the conditional mean value of  $Y$  for a given value of  $X$ . Since  $\beta_0 + \beta_1 X_i$  is the mean value, the sum of errors  $\sum_{i=1}^n \varepsilon_i = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)$  will be zero. In ordinary least squares, the objective is find the optimal values of  $\beta_0$  and  $\beta_1$  that will minimize the **Sum of Squares Errors** (SSE) given in Eq. (9.8):

$$SSE = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2 \quad (9.8)$$

To find the optimal values of  $\beta_0$  and  $\beta_1$  that will minimize SSE, we have to equate the partial derivative of SSE with respect to  $\beta_0$  and  $\beta_1$  to zero [Eqs. (9.9) and (9.11)].

$$\frac{\partial SSE}{\partial \beta_0} = \sum_{i=1}^n -2(Y_i - \beta_0 - \beta_1 X_i) = 2 \left( n\beta_0 + \beta_1 \sum_{i=1}^n X_i - \sum_{i=1}^n Y_i \right) = 0 \quad (9.9)$$

Solving Eq. (9.9) for  $\beta_0$ , the estimated value of  $\beta_0$  is given by

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \quad (9.10)$$

Differentiating SSE with respect to  $\beta_1$ , we get

$$\frac{\partial SSE}{\partial \beta_1} = \sum_{i=1}^n -2X_i(Y_i - \beta_0 - \beta_1 X_i) = -2 \sum_{i=1}^n (X_i Y_i - \beta_0 X_i - \beta_1 X_i^2) = 0 \quad (9.11)$$

Substituting the value of  $\beta_0$  from Eq. (9.10) in Eq. (9.11), we get

$$\frac{\partial SSE}{\partial \beta_1} = -2 \sum_{i=1}^n (X_i Y_i - X_i \bar{Y} + \beta_1 X_i \bar{X} - \beta_1 X_i^2) = \sum_{i=1}^n (X_i Y_i - X_i \bar{Y}) - \beta_1 \sum_{i=1}^n (X_i^2 - X_i \bar{X}) = 0$$

Thus, the value of  $\beta_1$  is given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i Y_i - X_i \bar{Y})}{\sum_{i=1}^n (X_i^2 - X_i \bar{X})} = \frac{\sum_{i=1}^n X_i (Y_i - \bar{Y})}{\sum_{i=1}^n X_i (X_i - \bar{X})} \quad (9.12)$$

Since  $\sum_{i=1}^n (\bar{X}\bar{Y} - Y_i \bar{X}) = 0$  and  $\sum_{i=1}^n [(\bar{X})^2 - X_i \bar{X}] = 0$ , the Eq. (9.12) can be re-written as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i Y_i - X_i \bar{Y}) + \sum_{i=1}^n (\bar{X}\bar{Y} - Y_i \bar{X})}{\sum_{i=1}^n (X_i^2 - X_i \bar{X}) + \sum_{i=1}^n [(\bar{X})^2 - X_i \bar{X}]} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \quad (9.13)$$

Since the **Pearson Correlation Coefficient**  $r = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y}$ , the estimated value of  $\beta_1$  can also be written as follows:

$$\hat{\beta}_1 = r \times \frac{\sigma_y}{\sigma_x} \quad (9.14)$$

where  $\bar{X}$  and  $\bar{Y}$  are the mean values of  $X$  and  $Y$  and  $\sigma_x$  and  $\sigma_y$  are the corresponding standard deviations.

### IMPORTANT

*The primary reason for using Sum of Squares Errors (SSE) for estimating the values of  $\beta_0$  and  $\beta_1$  is that minimizing SSE gives the best linear unbiased estimate.*

### EXAMPLE 9.1

#### Salary of Graduating MBA Students versus Their Percentage Marks in Grade 10

Table 9.2 provides the annual salary in rupees of 50 graduating MBA students of a Business School in 2016 and their corresponding percentage marks in grade 10 (File: Example 9.1.xlsx). Develop a simple linear regression model by estimating the model parameters.

**TABLE 9.2** Salary of MBA students versus their grade 10 marks

S. No.	Percentage in Grade 10	Salary	S. No.	Percentage in Grade 10	Salary
1	62.00	270000	26	64.60	250000
2	76.33	200000	27	50.00	180000
3	72.00	240000	28	74.00	218000
4	60.00	250000	29	58.00	360000
5	61.00	180000	30	67.00	150000
6	55.00	300000	31	75.00	250000
7	70.00	260000	32	60.00	200000
8	68.00	235000	33	55.00	300000
9	82.80	425000	34	78.00	330000
10	59.00	240000	35	50.08	265000
11	58.00	250000	36	56.00	340000
12	60.00	180000	37	68.00	177600
13	66.00	428000	38	52.00	236000
14	83.00	450000	39	54.00	265000
15	68.00	300000	40	52.00	200000
16	37.33	240000	41	76.00	393000
17	79.00	252000	42	64.80	360000
18	68.40	280000	43	74.40	300000

**TABLE 9.2** Salary of MBA students versus their grade 10 marks—Continued

S. No.	Percentage in Grade 10	Salary	S. No.	Percentage in Grade 10	Salary
19	70.00	231000	44	74.50	250000
20	59.00	224000	45	73.50	360000
21	63.00	120000	46	57.58	180000
22	50.00	260000	47	68.00	180000
23	69.00	300000	48	69.00	270000
24	52.00	120000	49	66.00	240000
25	49.00	120000	50	60.80	300000

Using Eqs. (9.10) and (9.13), the estimated values of  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are given by

$$\hat{\beta}_0 = 61555.3553 \text{ and } \hat{\beta}_1 = 3076.1774$$

The corresponding regression equation is given by

$$\hat{Y}_i = 61555.3553 + 3076.1774X_i$$

where  $\hat{Y}_i$  is the predicted value of  $Y$  for a given value of  $X_i$ .

The equation can be interpreted as follows: for every one percentage increase in grade 10 marks, the salary of the MBA students will increase at the rate of 3076.1774 on an average. The notations  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are used to denote that these are estimated values of the regression coefficients from the sample of 50 students. The Microsoft Excel output for SLR model is shown in Table 9.3.

**TABLE 9.3** Regression coefficient estimates using Microsoft Excel

	Coefficients	Standard Error	t-stat	p-value
Intercept	61555.35534	66701.901	0.9228	0.3607
Percentage in grade 10	3076.177438	1031.5258	2.9821	0.0044

### EXAMPLE 9.2

#### Healthcare Treatment Cost versus Age

Table 9.4 provides the treatment cost (in rupees) of 30 patients admitted at the ‘Die Another Day (DAD)’ hospital for cardiac ailments and their age in years (File name: Example 9.2.xlsx). Estimate the regression parameters.

**TABLE 9.4** Age versus treatment cost (in rupees)

S. No.	Age	Cost of Treatment	S. No.	Age	Cost of Treatment
1	22.0	125966	16	32.0	112346
2	26.0	128045	17	18.0	121345
3	26.5	128104	18	45.0	113290
4	27.7	128584	19	61.0	141785
5	30.0	128699	20	26.0	125098
6	31.2	130443	21	71.0	152198
7	8.1	120773	22	7.0	130484
8	9.0	121293	23	24.0	131346
9	10.0	121981	24	18.0	133026
10	12.0	122339	25	45.0	134243
11	7.0	122550	26	39.0	134648
12	14.0	123472	27	57.0	135778
13	11.0	124223	28	40.0	137693
14	17.0	125592	29	51.0	139067
15	21.0	125683	30	54.0	145559

Using Eqs. (9.10) and (9.13), the estimated values of  $\beta_0$  and  $\beta_1$  are given by

$$\hat{\beta}_0 = 118857.35 \text{ and } \hat{\beta}_1 = 348.5559$$

That is, for every one-year increase in age, the cost of treatment increases by 348.5559 rupees on an average.

## 9.5 | INTERPRETATION OF SIMPLE LINEAR REGRESSION COEFFICIENTS

Interpretation of regression coefficients is important for understanding the relationship between the response variable and the explanatory variable and the impact of change in the values of explanatory variables on the response variable. The interpretation will depend on the functional form of the relationship between the response and the explanatory variables. In this section, the interpretation of parameters for frequently used SLR models is discussed.

### 9.5.1 | Interpretation of $\beta_0$ and $\beta_1$ in $Y = \beta_0 + \beta_1 X$

When the functional form is  $Y = \beta_0 + \beta_1 X$ , the value of  $\beta_0 = E(Y|X=0)$ .

$\beta_1 = \frac{\partial Y}{\partial X}$ , that is  $\beta_1$  is the change in the value of  $Y$  for the unit change in the value of  $X$ , where  $\frac{\partial Y}{\partial X}$  is the partial derivative of  $Y$  with respect to  $X$ .

### 9.5.2 | Interpretation of $\beta_0$ and $\beta_1$ in $Y = \beta_0 + \beta_1 \ln(X)$

$$\beta_0 = E(Y|\ln(X) = 0) \text{ or } \beta_0 = E(Y|X = 1)$$

$$\frac{\partial Y}{\partial X} = \frac{1}{X} \beta_1 \Rightarrow \beta_1 = X \frac{\partial Y}{\partial X}$$

The above equation can be written as

$$\beta_1 = \frac{\partial Y}{(\partial X/X)} \Rightarrow \beta_1 \text{ is the change in } Y \text{ for percentage change in } X.$$

If one is interested in interpreting  $\beta_1$  for absolute change in  $X$ , then we can use the following equation:

$$E(Y|X+1) - E(Y|X) = \beta_1 \ln(X+1) - \beta_1 \ln(X) = \beta_1 \ln\left(\frac{X+1}{X}\right) \quad (9.15)$$

That is, for every one unit increase in  $X$ , the value of  $Y$  changes by  $\beta_1 \ln\left(\frac{X+1}{X}\right)$ .

### 9.5.3 | Interpretation of $\beta_0$ and $\beta_1$ in $\ln(Y) = \beta_0 + \beta_1 X$

$$\beta_0 = E(\ln(Y)|X = 0)$$

Differentiating the equation with respect to  $X$ , we get

$$\frac{1}{Y} \frac{\partial Y}{\partial X} = \beta_1 \Rightarrow \frac{\partial Y/Y}{\partial X} = \beta_1$$

$\beta_1$  is the percentage change in  $Y$  for a unit change in the  $X$  value. An alternative interpretation can be derived as follows:

$$Y = e^{(\beta_0 + \beta_1 X)}$$

Change in  $Y$  for one unit change in  $X$  is given by

$$e^{(\beta_0 + \beta_1 (X+1))} - e^{(\beta_0 + \beta_1 X)} = e^{(\beta_0 + \beta_1 X)} \times (e^{\beta_1} - 1) \quad (9.16)$$

That is, one unit change in  $X$  changes  $Y$  by a factor of  $e^{(\beta_0 + \beta_1 X)} \times (e^{\beta_1} - 1)$ .

### 9.5.4 | Interpretation of $\beta_0$ and $\beta_1$ in $\ln(Y) = \beta_0 + \beta_1 \ln(X)$

Differentiating the equation with respect to  $X$ , we get

$$\frac{1}{Y} \frac{\partial Y}{\partial X} = \frac{\beta_1}{X} \Rightarrow \beta_1 = \frac{\partial Y/Y}{\partial X/X}$$

$\beta_1$  is the percentage change in  $Y$  for a percentage change in  $X$ .



*The regression parameter values are valid only for the range of independent variable used in model building and it cannot be extrapolated beyond the range of data used in model building.*

## 9.6 | VALIDATION OF THE SIMPLE LINEAR REGRESSION MODEL

It is important to validate the regression model to ensure its validity and goodness of fit before it can be used for practical applications. The following measures are used to validate the simple linear regression models:

1. Co-efficient of determination ( $R$ -square).
2. Hypothesis test for the regression coefficient  $\beta_1$ .
3. Analysis of Variance for overall model validity (relevant more for multiple linear regression).
4. Residual analysis to validate the regression model assumptions.
5. Outlier analysis.

The above measures and tests are essential, but not exhaustive.

### 9.6.1 | Coefficient of Determination ( $R$ -Square or $R^2$ )

The primary objective of regression is to explain the variation in  $Y$  using the knowledge of  $X$ . The co-efficient of determination (or  $R$ -square or  $R^2$ ) measures the percentage of variation in  $Y$  explained by the model ( $\beta_0 + \beta_1 X$ ).

The simple linear regression model can be broken into explained variation and unexplained variation as shown in Eq. (9.17):

$$\underbrace{Y_i}_{\text{Variation in } Y} = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Variation in } Y \text{ explained by the model}} + \underbrace{\varepsilon_i}_{\text{Variation in } Y \text{ not explained by the model}} \quad (9.17)$$

In absence of the predictive model for  $Y_i$ , the users will use the mean value of  $Y_i$ . Thus, the total variation is measured as the difference between  $Y_i$  and  $\bar{Y}$  (i.e.,  $Y_i - \bar{Y}$ ). The total variation can be broken into components as shown in Table 9.5.

**TABLE 9.5** Description of total variation, explained variation and unexplained variation

Variation Type	Measure	Description
Total variation	$(Y_i - \bar{Y})$	Total variation is the difference between the actual value and the mean value.
Variation explained by the model	$(\hat{Y}_i - \bar{Y})$	Variation explained by the model is the difference between the estimated value of $Y_i$ and the mean value of $Y$
Variation not explained by model	$(Y_i - \hat{Y}_i)$	Variation not explained by the model is the difference between the actual value and the predicted value of $Y_i$ (error in prediction)

The relationship between the total variation, explained variation and the unexplained variation is given as follows:

$$\underbrace{Y_i - \bar{Y}}_{\text{Total Variation in } Y} = \underbrace{\hat{Y}_i - \bar{Y}}_{\text{Variation in } Y \text{ explained by the model}} + \underbrace{Y_i - \hat{Y}_i}_{\text{Variation in } Y \text{ not explained by the model}}$$

It can be proved mathematically that sum of squares of total variation is equal to sum of squares of explained variation plus sum of squares of unexplained variation (for proof refer to Kutner *et al.*, 2013, page 65):

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (9.18)$$

where  $SST$  is the sum of squares of total variation,  $SSR$  is the sum of squares of variation explained by the regression model and  $SSE$  is the sum of squares of errors or unexplained variation. The coefficient of determination ( $R^2$ ) is given by

$$\text{Coefficient of determination} = R^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{SSR}{SST} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \quad (9.19)$$

Since  $SSR = SST - SSE$ , Eq. (9.19) can be written as

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^n (\hat{Y}_i - Y_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \quad (9.20)$$

Thus,  $R^2$  is the proportion of variation in response variable  $Y$  explained by the regression model. Coefficient of determination ( $R^2$ ) has the following properties:

1. The value of  $R^2$  lies between 0 and 1.
2. Higher value of  $R^2$  implies better fit, but one should be aware of spurious regression.
3. Mathematically, the square of correlation coefficient is equal to coefficient of determination (i.e.,  $r^2 = R^2$ ).
4. We do not put any minimum threshold for  $R^2$ ; higher value of  $R^2$  implies better fit. However, a minimum value of  $R^2$  for a given significance value  $\alpha$  can be derived using the relationship between the F-statistic and  $R^2$  (discussed in Section 9.6.4).



*Coefficient of Determination is the proportion (or percentage) of variation in  $Y$  as explained by the regression model.*

The simple linear regression model output for Example 9.1 using regression function in Microsoft Excel is shown in Table 9.6.

In Table 9.6, the  $R$ -square value is 0.1563, that is, grade 10 marks explain 15.63% of the variation in the starting salary of MBA students.

**TABLE 9.6** Regression output for Example 9.1 (using Excel regression function)

Regression Statistics					
Multiple $R$	0.39536731				
$R$ Square	0.15631531 10 <sup>th</sup> grade marks explain 15.63% variation in starting salary of the MBA students.				
Adjusted $R$ Square	0.13873854				
Standard Error	71195.4556				
Observations	50				
ANOVA					
	$df$	$SS$	$MS$	$F$	Significance $F$
Regression	1	4.51E+10	4.51E+10	8.893293	0.004487
Residual	48	2.43E+11	5.07E+09		
Total	49	2.88E+11			
Regression Coefficient					
	Coefficients	Standard Error	$t$ -stat	$p$ -value	
Intercept	61555.3553	66701.9	0.922843	0.360705	
Percent_Grade 10	3076.17744	1031.526	2.982162	0.004487	

### 9.6.2 | Spurious Regression

One of the major problems with coefficient of determination ( $R^2$ ) is that two sets of data without any relationship can have a very high coefficient of determination value. The data in Table 9.7 shows the number of Facebook users (in millions) and the number of people who died of helium poisoning in UK between 2004 and 2012. Table 9.8 provides the regression output from Microsoft Excel for the data in Table 9.7.

**TABLE 9.7** Number of Facebook users and the number of people who died of helium poisoning in UK

Year	Number of Facebook users in millions ( $X$ )	Number of people who died of helium poisoning in UK ( $Y$ )
2004	1	2
2005	6	2
2006	12	2
2007	58	2
2008	145	11
2009	360	21
2010	608	31
2011	845	40
2012	1056	51

**TABLE 9.8** Facebook users versus helium poisoning in UK regression output

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.996442				
R Square	0.992896				
Standard Error	1.69286				
Observations	9				
ANOVA					
	df	SS	MS	F	Significance F
Regression	1	2803.94	2803.94	978.4229	8.82E-09
Residual	7	20.06042	2.865775		
Total	8	2824			
	Coefficients	Standard Error	t-stat	p-value	Lower 95%
Intercept	1.9967	0.76169	2.62143	0.034338	0.195607
FB	0.0465	0.00149	31.27975	8.82E-09	0.043074
					0.050119

The  $R$ -square value (in Table 9.8) for regression model between the number of deaths due to helium poisoning in UK and the number of Facebook users is 0.9928. That is, 99.28% variation in the number of deaths due to helium poisoning in UK is explained by the number of Facebook users. The regression model is given as follows:

$$Y = 1.9967 + 0.0465 X$$

That is, for every 100 million increase in Facebook users about 4.6 more people are dying in the UK due to helium poisoning (well I always knew that Facebook can cause a few social problems!). Such spurious regressions are a major concern in regression model. Thus, one has to be careful about using regression model purely based on  $R$ -square value.



A high  $R$ -square value is not necessarily a good indicator of the correctness of the model; it could be a spurious relationship.

### 9.6.3 | Hypothesis Test for Regression Co-Efficient (t-Test)

The regression co-efficient ( $\beta_1$ ) captures the existence of a linear relationship between the response variable and the explanatory variable. If  $\beta_1 = 0$ , we can conclude that there is no statistically significant linear relationship between the two variables. In this section, we will discuss the appropriate test statistic to carry out a hypothesis test on the regression coefficient  $\beta_1$ .

The estimate of  $\beta_1$  using OLS is given by (Kutner *et al.*, 2013, page 18)

$$\beta_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2} - \frac{\bar{X}\sum_{i=1}^n (Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (9.21)$$

Equation (9.21) can be written as follows:

$$\beta_1 = \frac{\sum_{i=1}^n K_i Y_i}{\sum_{i=1}^n K_i^2} \text{ where } K_i = (X_i - \bar{X}) \quad (9.22)$$

That is, the value of  $\beta_1$  is a function of  $Y_i$  ( $K_i$  is a constant since  $X_i$  is assumed to be non-stochastic). Since we assume that the error terms follow a normal distribution, the  $Y_i$  values also follow normal distribution. Since the value of  $\beta_1$  is a constant multiple of  $Y_i$ , it follows a normal distribution. Note that  $\beta_1$  is estimated using a sample data set from a population that follows a normal distribution. Thus the sampling distribution of  $\beta_1$  is a  $t$ -distribution with  $(n - 2)$  degrees of freedom, since the standard error (standard deviation estimated from a sample) of  $\beta_1$  is estimated from the sample itself after estimating two parameters  $\beta_0$  and  $\beta_1$ . The standard error of  $\beta_1$  is given by

$$S_e(\hat{\beta}_1) = \frac{S_e}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}} \quad (9.23)$$

In Eq. (9.23),  $S_e$  is the standard error of estimate (or standard error of the residuals) that measures the accuracy of prediction and is given by

$$S_e = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^n \epsilon_i^2}{n-2}} \quad (9.24)$$

The denominator in Eq. (9.24) is  $(n - 2)$  since  $\beta_0$  and  $\beta_1$  are estimated from the sample in estimating  $Y_i$  and thus two degrees of freedom are lost. The standard error of  $\hat{\beta}_1$  can be written as

$$S_e(\hat{\beta}_1) = \frac{S_e}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}} = \frac{\sqrt{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 / n-2}}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}} \quad (9.25)$$



Since  $\beta_1 = \sum_i K_i Y_i / \sum_i K_i^2$ , and  $Y_i$  is assumed to follow a normal distribution,  $t$ -test with  $(n - 2)$  df is used for checking whether  $\beta_1$  is zero or not.

The null and alternative hypotheses for the SLR model can be stated as follows:

$$\begin{aligned} H_0 &: \text{There is no relationship between } X \text{ and } Y \\ H_A &: \text{There is a relationship between } X \text{ and } Y \end{aligned}$$

$\beta_1 = 0$  would imply that there is no linear relationship between the response variable  $Y$  and the explanatory variable  $X$ . Thus, the null and alternative hypotheses can be restated as follows:

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_A &: \beta_1 \neq 0 \end{aligned}$$

The corresponding  $t$ -statistic is given as follows:

$$t = \frac{\hat{\beta}_1 - \beta_1}{S_e(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - 0}{S_e(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{S_e(\hat{\beta}_1)} \quad (9.26)$$

In Example 9.1, the  $t$ -value for the variable percentage marks grade 10 is given by

$$t = \frac{\hat{\beta}_1}{S_e(\hat{\beta}_1)} = \frac{3076.1774}{1031.526} = 2.9821$$

The corresponding **degrees of freedom** ( $df$ ) is  $(n - 2)$  which in this case is  $50 - 2 = 48$ . Here two degrees of freedom are lost since  $\beta_0$  and  $\beta_1$  are estimated from the data. This is a two-tailed test, the critical  $t$ -value is 2.01 for  $\alpha = 0.05$  and  $df = 48$ . The  $p$ -value corresponding to  $t = 2.9821$  with 48 degrees of freedom is 0.0044 (Table 9.6). Since the  $p$ -value is less than 0.05, we reject the null hypothesis and conclude that there is significant evidence suggesting a linear relationship between  $X$  and  $Y$ .

#### 9.6.4 | Test for Overall Model: Analysis of Variance (F-test)

Using the **Analysis of Variance** (ANOVA), we can test whether the overall model is statistically significant. However, for a simple linear regression, the null and alternative hypotheses in ANOVA and  $t$ -test are exactly same and thus there will be no difference in the  $p$ -value.

The null and alternative hypothesis for  $F$ -test are given by

$H_0$ : There is no statistically significant relationship between  $Y$  and any of the explanatory variables (i.e., all regression coefficients are zero).

$H_A$ : Not all regression coefficients are zero.

Alternatively:

$H_0$ : All regression coefficients are equal to zero.

$H_A$ : Not all regression coefficients are equal to zero.

The  $F$ -statistic is given by

$$F = \frac{MSR}{MSE} = \frac{SSR / 1}{SSE / n - 2} \quad (9.27)$$

For Example 9.1, the  $F$ -value is given by

$$F = \frac{MSR}{MSE} = \frac{SSR / 1}{SSE / n - 2} = \frac{4.51 \times 10^{10}}{5.07 \times 10^9} = 8.8932$$

In Table 9.6, the  $p$ -value corresponding to  $F$ -statistic value of 8.8932 is 0.0044. Since the  $p$ -value is less than 0.05 (assume that  $\alpha = 0.05$ ), the null hypothesis is rejected. Note that the  $p$ -value of  $t$ -test and  $F$ -test are same in Table 9.6. This is due to the fact that the model has only one independent variable and the null hypothesis for both  $t$ -test and  $F$ -test are identical (in SLR,  $F = t^2$ ). The mathematical relationship between  $F$ -statistic and  $R^2$  in a simple linear regression is given by

$$F = \frac{R^2}{(1-R)^2 / (n-2)} \quad (9.28)$$

The above relationship can be used for calculating the minimum value  $R^2$  required for a statistically significant relationship between two variables. In Table 9.6, the value of  $R^2$  is 0.156315. Using the formula in Eq. (9.18), the  $F$ -statistic value is as follows:

$$F = \frac{R^2}{(1-R)^2 / (n-2)} = \frac{0.156315}{(1-0.156315) / 48} = 8.8932$$



*In a simple linear regression  $F = t^2$  and the  $p$ -value is same since  $F$ -test and  $t$ -test are equivalent for SLR.*

### 9.6.5 | Residual Analysis

Residual (error) analysis is important to check whether the assumptions of regression models have been satisfied. It is performed to check the following:

1. The residuals ( $Y_i - \hat{Y}_i$ ) are normally distributed.
2. The variance of residual is constant (homoscedasticity).
3. The functional form of regression is correctly specified.
4. If there are any outliers.

#### **Checking for Normal Distribution of Residuals ( $Y_i - \hat{Y}_i$ )**

Although for OLS estimation we do not need to assume that the residuals follow normal distribution, it is an important assumption for the hypothesis tests [refer to Eqs. (9.21), (9.26) and (9.27)]. The easiest technique to check whether the residuals follow normal distribution is to use the P-P plot (Probability-Probability plot). The P-P plot compares the cumulative distribution function of two probability distributions against each other. In the current context, we use the P-P plot to check whether the distribution

of the residual matches with that of a normal distribution. The P-P plots (using software SPSS) for the regression models developed in Examples 9.1 and 9.2 are shown in Figures 9.5 and 9.6, respectively.

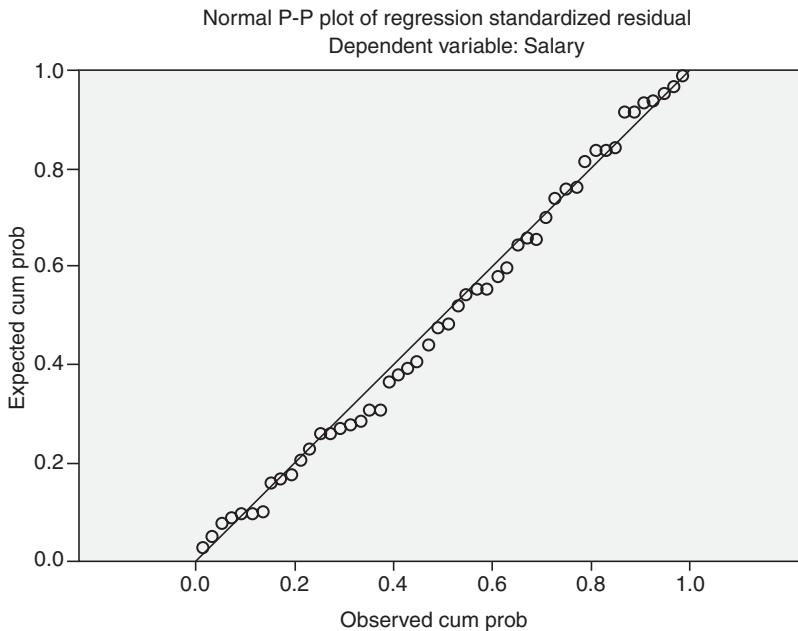


FIGURE 9.5 Residual plot (P - P Plot) of the regression model  $Y = \beta_0 + \beta_1 \text{salary}$  (Example 9.1).

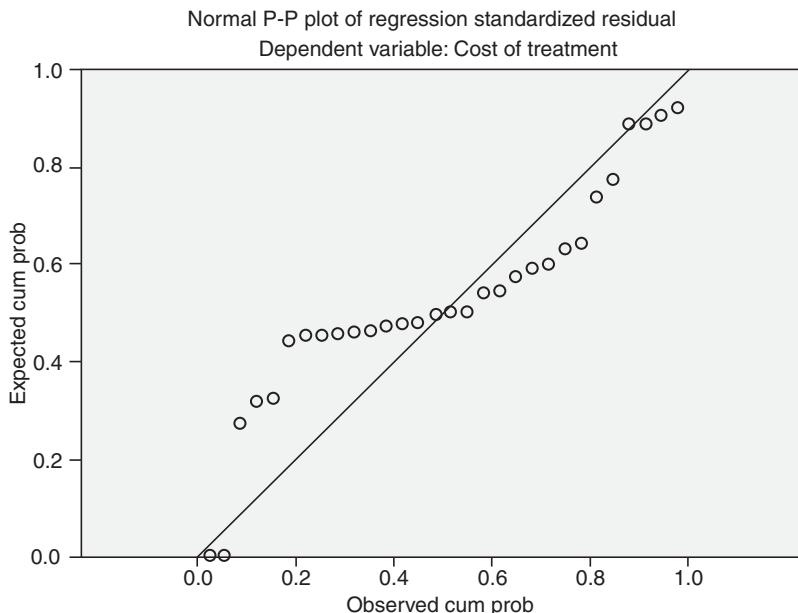
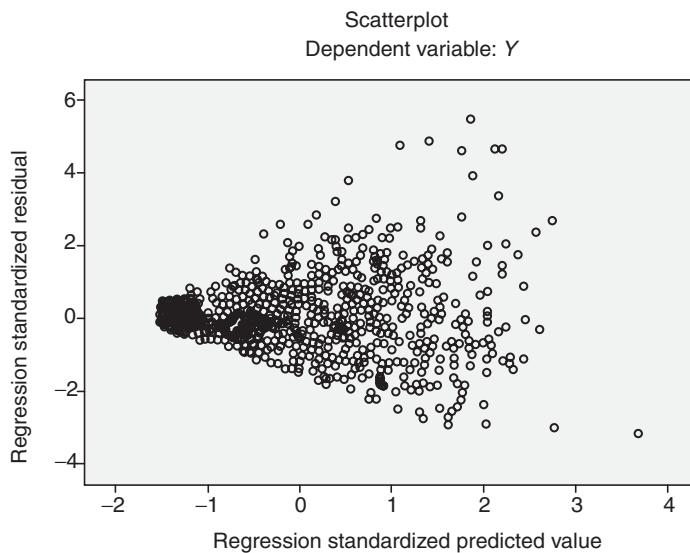


FIGURE 9.6 Residual plot (P - P Plot) of the regression model  $Y = \beta_0 + \beta_1 \text{age}$  (Example 9.2).

In Figure 9.5, the diagonal line is the cumulative distribution of a normal distribution, whereas the dots represent the cumulative distribution of the residuals. Since the dots are close to the diagonal line, we can conclude that the residuals follow an approximate normal distribution (we need only an approximate normal distribution). Thus, the hypothesis tests ( $t$  and  $F$ ) are valid for the data in Example 9.1. However, the residual plot in Figure 9.6 does not follow normal distribution (since the dots are away from the diagonal line), which puts doubt about the validity of the model itself, since we cannot trust the outcome of hypothesis tests (the  $p$ -value) of Example 9.2.

### Test of Homoscedasticity

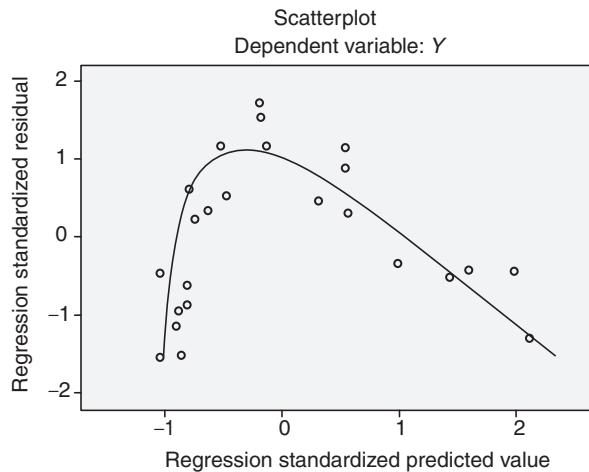
An important assumption of regression model is that the residuals have constant variance (homoscedasticity) across different values of the explanatory variable ( $X$ ). That is, the variance of residuals is assumed to be independent of variable  $X$ . Failure to meet this assumption will result in unreliability of the hypothesis tests. Figure 9.7 is the plot of standardized predicted values versus the standardized residuals. If there is heteroscedasticity (non-constant variance of residuals) then we can expect a funnel type shape in the residual plot (as shown in Figure 9.7). A funnel shape indicates that the variance of residuals depends on the value of independent variable  $X$ .



**FIGURE 9.7** Funnel shape in the standardized residual plot indicates heteroscedasticity.

### Testing the Functional Form of Regression Model

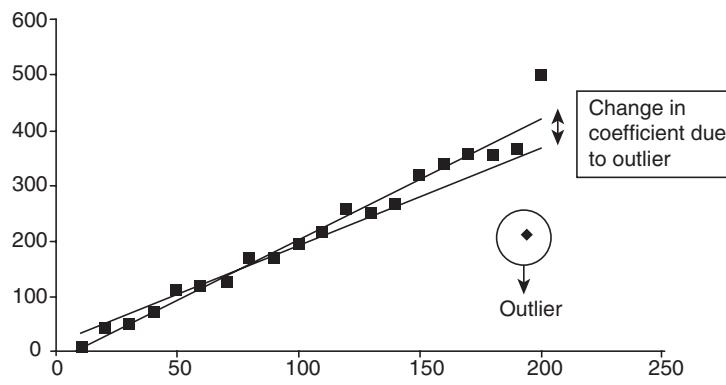
Any pattern in the residual plot would indicate incorrect specification (misspecification) of the model. Residual plot in Figure 9.8 shows a parabolic shape indicating the model mis-specification, that is, an incorrect functional form is used.



**FIGURE 9.8** A pattern (parabola) in the residual plot indicates model misspecification.

## 9.7 | OUTLIER ANALYSIS

Outliers are observations whose values show a large deviation from mean value, that is  $(Y_i - \bar{Y})$  is large. Presence of an outlier can have significant influence on values of regression coefficients. Thus, it is important to identify the existence of outliers in the data. Figure 9.9 shows change in the value of regression coefficient due the presence of outliers.



**FIGURE 9.9** Influence of outliers on regression coefficients.

The following distance measures are useful in identifying the influential observations:

1. Z-Score
2. Mahalanobis Distance
3. Cook's Distance
4. Leverage Values
5. DFBeta and DFFit values

We will discuss them in the following subsections.

### 9.7.1 | Z-Score

Z-score is the standardized distance of an observation from its mean value. For the predicted value of the dependent variable  $Y$ , the Z-score is given by

$$Z = \left( \frac{\hat{Y}_i - \bar{Y}}{\sigma_Y} \right) \quad (9.29)$$

where  $\bar{Y}$  and  $\sigma_Y$  are, respectively, the mean and the standard deviation of dependent variable estimated from the sample data. Any observation with a Z-score of more than 3 may be flagged as outlier and influential observations that may change the regression parameter values significantly.

### 9.7.2 | Mahalanobis Distance

Mahalanobis distance is the distance between specific values of the independent variable ( $X_i$ ) to the centroid of all observations of the explanatory variable. Mahalanobis distance value of more than chi-square critical value (with degrees of freedom is equal to the number of explanatory variables) is classified as outliers.

### 9.7.3 | Cook's Distance

Cook's distance measures how much the predicted value of the dependent variable changes for all the observations in the sample when a particular observation is excluded from sample for the estimation of regression parameters. Cook's distance for simple linear regression is given by (Ryan, 2009, Kutner et al 2013)

$$D_i = \frac{\sum_j (\hat{Y}_j - \hat{Y}_{j(i)})^2}{(k+1) \times MSE} \quad (9.30)$$

where  $D_i$  is the Cook's distance measure for  $i^{\text{th}}$  observation,  $k$  is the number of predictors in the model,  $\hat{Y}_j$  is the predicted value of  $j^{\text{th}}$  observation including  $i^{\text{th}}$  observation,  $\hat{Y}_{j(i)}$  is the predicted value of  $j^{\text{th}}$  observation after excluding  $i^{\text{th}}$  observation from the sample, MSE is the Mean-Squared-Error. A Cook's distance value of more than 1 indicates highly influential observation.

### 9.7.4 | Leverage Value

Leverage value of an observation measures the influence of that observation on the overall fit of the regression function. Leverage value for an observation in SLR is given by (Ryan, 2009)

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (9.31)$$

Leverage value of more than  $2/n$  or  $3/n$  is treated as highly influential observation. In Eq. (9.31), the first term ( $1/n$ ) will tend to zero for large value of  $n$ . The Mahalanobis distance, Cook's distance and Leverage values for the first 10 cases in Example 9.1 is shown in Table 9.9.

**TABLE 9.9** Distance measures for first 10 cases in Table 9.2

S. No.	Percentage in Grade10	Salary	Mahalanobis Distance	Cook's Distance	Leverage Value
1.0	62.00	270,000	0.03801	0.00067	0.00078
2.0	76.33	200,000	1.58353	0.05336	0.03232
3.0	72.00	240,000	0.67115	0.00659	0.01370
4.0	60.00	250,000	0.15825	0.00004	0.00323
5.0	61.00	180,000	0.08785	0.01076	0.00179
6.0	55.00	300,000	0.81887	0.01872	0.01671
7.0	70.00	260,000	0.37994	0.00083	0.00775
8.0	68.00	235,000	0.17103	0.00310	0.00349
9.0	82.80	425,000	3.66560	0.13495	0.07481
10.0	59.00	240,000.0	0.24923	0.00002	0.00509

### 9.7.5 | DFFit and DFBeta

DFFit is the change in the predicted value of  $Y_i$  when case  $i$  is removed from the data set. DFBeta is the change in the regression coefficient values when an observation  $i$  is removed from the data. DFFit and DFBeta will be discussed in detail in Chapter 10.

Once an influential observation is found in the data, an important question that arises is how to handle an outlier or influential observation. We will discuss this issue in detail in Chapter 10.

## 9.8 | CONFIDENCE INTERVAL FOR REGRESSION COEFFICIENTS $\beta_0$ AND $\beta_1$

The points estimates of the regression parameters  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are obtained using OLS estimation and are given by Eqs. (9.10) and (9.12). The standard error of estimates of  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are given by

$$S_e(\hat{\beta}_0) = \frac{S_e \times \sqrt{\sum_{i=1}^n X_i^2}}{\sqrt{n \times SS_X}} \quad (9.32)$$

$$S_e(\hat{\beta}_1) = \frac{S_e}{\sqrt{SS_X}} \quad (9.33)$$

where

$$S_e = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}} \quad (9.34)$$

where  $S_e$  is the standard error of residuals and  $SS_X = \sum_{i=1}^n (X_i - \bar{X})^2$ .

The interval estimate or  $(1-\alpha)100\%$  confidence interval for  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are given by

$$\hat{\beta}_0 \mp t_{\alpha/2,n-2} S_e (\hat{\beta}_0) \quad (9.35)$$

$$\hat{\beta}_1 \mp t_{\alpha/2,n-2} S_e (\hat{\beta}_1) \quad (9.36)$$

The confidence interval for  $\hat{\beta}_0$  and  $\hat{\beta}_1$  in Example 9.1 is shown in Table 9.10. The 95% confidence interval for  $\hat{\beta}_0$  is  $(-72557.805, 195668.515)$  and the 95% confidence interval for  $\hat{\beta}_1$  is  $(1002.156, 5150.199)$ .

**TABLE 9.10** Confidence interval for  $\hat{\beta}_0$  and  $\hat{\beta}_1$  for Example 9.1

Model	Unstandardized Coefficients		t	Sig.	95.0% Confidence Interval for $B$	
	B	Std. Error			Lower Bound	Upper Bound
1	(Constant) 61555.355	66701.901	.923	.361	-72557.805	195668.515
	Percentage in grade 10 3076.177	1031.526	2.982	.004	1002.156	5150.199

## 9.9 | CONFIDENCE INTERVAL FOR THE EXPECTED VALUE OF Y FOR A GIVEN X

The regression model  $Y_i = \beta_0 + \beta_1 X_i$  gives the conditional expected value of  $Y_i$  for a given value of  $X_i$ . The SLR model gives us the point estimate of the conditional expected value (or mean value) of the dependent variable for a given value of the independent variable. Since the point estimates are subjected to higher levels of error, due to uncertainties around estimation of parameters and natural variation in the data around the predicted line, the user would like to know the interval estimate or the **confidence interval** for the conditional expected value. The confidence interval of the expected value of  $Y_i$  for a given value of  $X_i$  is given by

$$\hat{Y}_i \pm t_{\alpha/2,n-2} \times S_e \times \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}} \quad (9.37)$$

where the term  $S_e \times \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$  is the standard error of  $E(Y|X)$ .

For large  $n$ , the confidence interval of  $E(Y|X)$  will converge to  $\hat{Y}_i$ . This is because, as  $n \rightarrow \infty$ , the term  $\sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}} \rightarrow 0$

## 9.10 | PREDICTION INTERVAL FOR THE VALUE OF Y FOR A GIVEN X

In many applications, we would be interested in knowing the interval estimate of  $Y_i$  for a given value of  $X_i$  (called **prediction interval**). The prediction interval of  $Y_i$  for a given value of  $X_i$  is given by

$$\hat{Y}_i \pm t_{\alpha/2, n-2} \times S_e \times \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}} \quad (9.38)$$

where the term,  $S_e \times \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$  is the standard error of  $Y_i$  for a given  $X_i$  value.

For large  $n$ , the confidence interval of  $E(Y|X)$  will converge to

$$\hat{Y}_i \pm t_{\alpha/2, n-2} \times S_e$$

This is because, as  $n \rightarrow \infty$ , the term  $\sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$  converges to 1.



*Confidence interval is the interval estimate of conditional expected value of  $Y_i$ , that is  $E(Y_i|X_i)$ . Whereas prediction interval is the interval estimate for value of  $Y_i$  for a given  $X_i$ .*

### EXAMPLE 9.3

Using the data in Example 9.1 (a) calculate the confidence interval and prediction interval for the salary of a student with a score of 60% in grade 10. (b) What is the probability that his annual salary will be more than 3,00,000 per annum?

**Solution:**

- (a) We have to use Eqs. (9.37) and (9.38) to calculate the confidence interval and prediction interval, respectively:

$$\hat{Y}_i = 61555.3553 + 3076.1774 \times 60 = 246126.002$$

$$\bar{X} = 63.9224; (X_i - \bar{X})^2 = (60 - 63.9224)^2 = 15.3852; \sum_{i=1}^{50} (X_i - \bar{X})^2 = 4763.70$$

$$S_e = 71195.4556 \text{ and } t_{0.025, 48} = 2.0106$$

Substituting the values in Eq. (9.37), the 95% confidence interval is given by (2,24,308.4 – 2,67,943.6). The 95% prediction interval using Eq. (9.38) is given by (1,01,327.3 – 3,90,927.2)

- (b) Note that  $Y_i$  is a normal distribution, where the mean is 246126 for  $X_i = 60\%$  and the standard error is 71195.46. The corresponding probability that the salary will be more than 3,00,000 is 0.2246. That is, 22.46% of the students with 60% in grade 10 would earn a salary of more than INR 3,00,000 per annum.

Note we have to use  $S_e \times \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$  for the standard error for small sample.

### Package Pricing at the Die Another Day (DAD) Hospital

#### Case Study

Dr Ajit Narayanan was watching the fresh green paddy fields from his Tata Innova while driving on national highway 47 (NH 47) between Palakkad and Thrissur in the God's own country, Kerala. He turned towards Professor Dinesh Kumar and said:

There is a reason why this land is called the God's own country. I don't think there are many places in the world that can rival the beauty of this land; especially after monsoon.

Professor Dinesh Kumar nodded his head. Kerala was not new to him since he was born in Palakkad. However, his mind was preoccupied with the main reason for visiting Dr Ajit Narayanan and Thrissur. Dr Narayanan was the CEO of a multi-speciality hospital called Die Another Day (DAD) Hospital<sup>1</sup> in Thrissur. Although, the hospital treated patients of all illnesses, their speciality was cardiology and DAD was a very popular hospital among Keralites. On an average, they conducted 200 heart surgeries every month and were in the process of increasing the monthly capacity to 400 heart surgeries.

Dr Ajit mentioned the problem he was currently facing as the CEO of the DAD. He said:

"Professor, the business models in healthcare are changing fast. As a hospital, we publish the price list for most of the treatments. We also negotiate package prices with many state governments for specific ailments. These government schemes are created for the benefit of economically weaker sections of the society. The package pricing is also creating intense competition between the hospitals; there is a price war for each treatment. But we don't know whether this is the right strategy. Sometimes I feel that we should charge like the old days where the patient pays for all the costs associated with the treatment and the consultancy fee."

<sup>1</sup> Name changed to maintain confidentiality.

**Continued...**

**Case Study** Professor Dinesh Kumar was aware of this new business model that was becoming very popular among hospitals across India and other parts of the world. Many hospitals quoted package prices (flat rate) for treatments such as heart surgery, knee replacement, etc. Irrespective of the expenses and the duration of treatment, the patient would pay only the agreed price since it was a contract between the patient and the hospitals. Hospitals cannot charge more than the package price under any circumstance. Many state governments' insisted on such contracts as there was a perception in the public that the hospitals insisted on unnecessary diagnostic tests whose profit margin was high.

Dr Ajit Narayanan commented that there was a high risk involved with flat fee pricing since the actual cost could far exceed the package price. He said:

It is like buffet pricing in restaurants, in which a customer pays a fixed price and can eat as much as he can. Restaurants have control over the menu, but hospitals don't have any control over the number of days it will take to a cure a person because it depends heavily on an individual body!

There are many decisions that Dr Ajit has to take – whether to use package pricing or traditional pricing? Should package pricing be offered to all types of treatments if they plan to have package pricing strategy? How should one come up with the package pricing and how to use package pricing as a competitive strategy in the market since he was expecting many new hospitals to come up in Thrissur in the next couple of years.

#### CASE QUESTIONS

Use the data on body weight of patients and their treatment cost provided in the data file “DAD Hospital Data.xlsx” and answer the following questions:

1. Is there a statistical evidence to support that the cost of treatment and body weight are related? Support your answer with all necessary tests.
2. Comment on the value of  $R$ -square. Does a low  $R$ -square value indicate that the model is not useful?
3. Interpret the value of the coefficient of weight in the model developed in question 1. What will be average difference in cost of treatment for patient aged 50 and patient aged 51?
4. Is it possible to conclude that a patient weighing 50 kg is likely to spend at least INR 500 more than the one weighing 49 kg at 90% confidence level?
5. At the time of admission, a patient's body weight is 50 kg. At 95% confidence level, what will be the maximum cost of treatment for this patient?
6. For a patient weighing 50 kg, what is the 95% confidence interval for the average cost of treatment?
7. DAD hospital is planning to introduce package price for the treatment and they would like to charge INR 3,00,000 for the patients weighing 50 kg. That is, the patient is charged

**Continued...**

INR 3,00,000 irrespective of the actual treatment cost. What is the probability that the treatment cost is likely to exceed the package price?

**Suggested Answers to the Case Questions**

1. Is there a statistical evidence to support that the cost of treatment and the body weight are related? Support your answer with all necessary tests.

**Answer:** We have to develop a simple linear regression model and validate the model to check whether there is a linear relationship between the cost of treatment and the weight.

Let  $Y$  = cost of treatment and  $X$  = weight of the patient. The corresponding simple linear regression model is given by

$$Y = \beta_0 + \beta_1 \text{Body weight} \quad (9.39)$$

The data set ‘DAD Hospital Data.xls’ has cost of treatment and weight for 120 patients admitted to the DAD hospital. The regression output for the model using the software SPSS is shown in Tables 9.11 and 9.12.

**TABLE 9.11** The regression model summary

Model	<i>r</i> (correlation coefficient)	R-Square (coefficient of determination)	Std. Error of the Estimate
1	0.198	0.039	96522.63093

**TABLE 9.12** Regression coefficients

Model	Unstandardized Coefficients		Standardized Coefficients		<i>t</i>	Sig.
	<i>B</i>	Std. Error	Beta			
1	(Constant) 127498.079	43832.757			2.909	0.004
	Body Weight 1678.933	763.569	0.198			

That is, the relationship between the cost of treatment and the body weight is given by

$$Y = 127498.079 + 1678.933 \times \text{Body Weight}$$

The *p*-value for the coefficient “Body Weight” is 0.030 which is less than 0.05; thus, the independent variable body weight is significant at  $\alpha = 0.05$  or at 95% confidence level. From the model we can interpret that the cost of treatment increases at the rate of INR 1678.933 per 1 kg increase in the body weight. However, before we accept the model, we have to check the important assumptions of normality and homoscedasticity. Figure 9.10 is the P-P plot that shows the observed cumulative probability of standardized residuals and expected cumulative probability of a normal distribution (diagonal line). Figure 9.11 is a plot between the standardized residual and the standardized response

• Case Study

Continued...

variable ( $Y$ ). The plot between residual and independent variable values can also be used for finding existence of heteroscedasticity.

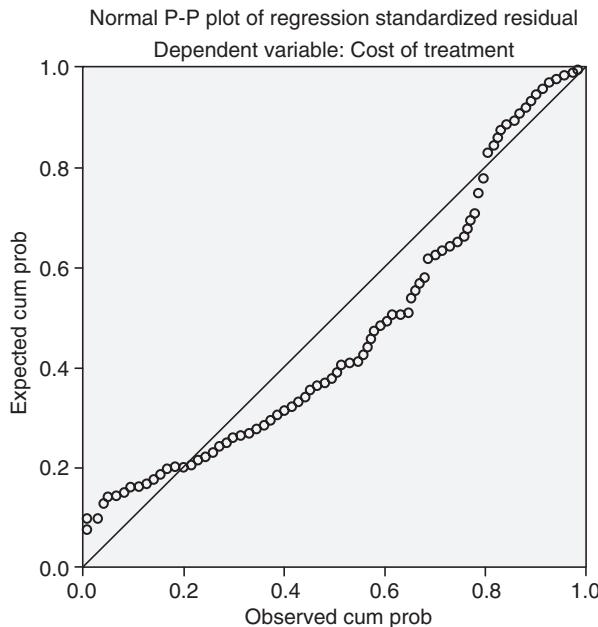


FIGURE 9.10 P-P plot for the model described in Eq. (9.28).

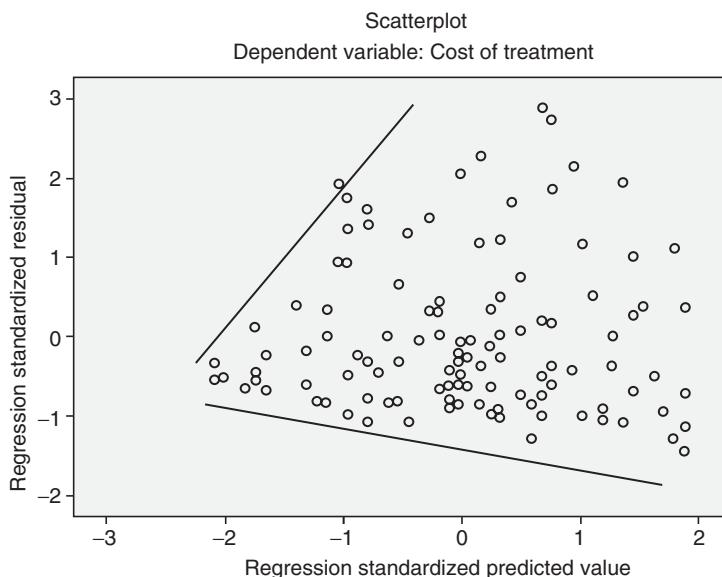


FIGURE 9.11 Plot of standardized predicted versus standardized residual for model.

**Continued...**

It is evident from Figures 9.10 and 9.11 that both the normality and homoscedasticity assumptions are not satisfied by the model defined in Eq. (9.39), which puts doubt over the model.

Whenever the assumptions of regression model are not met, we have to use a remedial measure and one of the popular remedial measures is **Transformation of Variables** (transformation of variables will be discussed in Chapter 10). In this case, we try the following model in which instead of  $Y$ , we build the model between  $\ln(Y)$  and  $X$ , where  $\ln(Y)$  is natural logarithm of  $Y$ :

$$\ln(Y) = \alpha_0 + \alpha_1 \times \text{Body Weight} \quad (9.40)$$

The model outputs for the regression Eq. (9.40) are provided in Tables 9.13 and 9.14.

**TABLE 9.13** The regression model summary

Model	R	R-Square	Std. Error of the Estimate
1	0.214	0.046	0.3975

**TABLE 9.14** The regression coefficient

Model	Unstandardized Coefficients		t-value	Sig.
	B	Std. Error		
1	(Constant)	11.804	0.181	65.381
	Body Weight	0.0074	0.003	2.377

That is, the relationship between the cost of treatment and the weight is given by

$$\ln(Y) = 11.804 + 0.0074 \times \text{Body Weight}$$

The  $p$ -value for the coefficient ‘body weight’ is less than 0.05, thus the variable body weight is significant at 95% confidence level. Figures 9.12 and 9.13 provide the P-P plot and the residual plot between the standardized residual and the standardized response variable  $\ln(Y)$ .

Figure 9.12 (for normality) and Figure 9.13 (for homoscedasticity) are looking better than Figures 9.10 and 9.11. Thus, the model in Eq. (9.40) may be used for predicting the cost of treatment since it satisfies important assumptions of SLR model.

- Comment on the value of the  $R$ -square. Does a low  $R$ -square value indicate that the model is not useful?

**Answer:** The  $R$ -square value for the model  $\ln(Y) = \alpha_0 + \alpha_1 \times \text{Body Weight}$  is only 0.046. That is, the model is explaining only 4.6% of the variation in the value of  $\ln(Y)$ . Low  $R$ -square values do not imply that the model is not useful. The primary objective of regression is to find whether there is a relationship between the response variable (cost of treatment) and the independent variable (body weight of the patient). The regression model establishes this relationship since the  $p$ -value of the weight coefficient is less than 0.05

Case Study •

Continued...

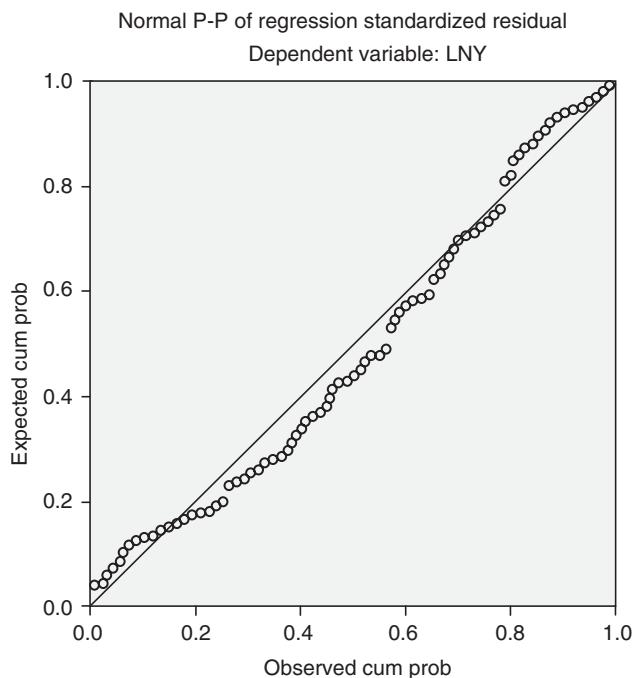


FIGURE 9.12 P-P plot for the model described in Eq. (9.29).

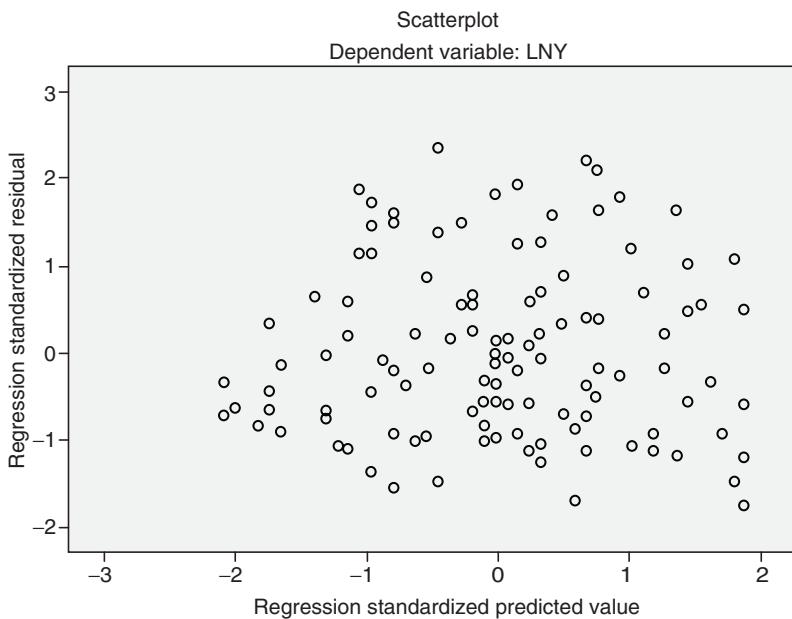


FIGURE 9.13 Plot of standardized predicted versus standardized residual.

**Continued...**

and both normality and homoscedasticity assumptions are satisfied reasonably. Low *R*-square may create problem when we use the model for prediction since the error is likely to be higher.

- Interpret the value of the coefficient of weight in the model developed in question 1. What will be average difference in cost of treatment for someone aged 50 and 51?

**Answer:** The regression model is given by

$$\begin{aligned}\ln(Y) &= 11.804 + 0.0074 \times \text{Body Weight} \\ \Rightarrow Y &= \exp(11.804 + 0.0074 \times \text{Body Weight})\end{aligned}$$

The coefficient for weight is 0.0074, that is, for every 1 kg increase in weight, the cost of treatment increases by a factor of  $e^{11.804+0.0074X} (e^{0.0074} - 1)$ . The average cost of treatment for persons aged 50 and 51 are given by:

$$\begin{aligned}X = 50; Y &= \exp(11.804 + 0.0074 \times 50) = \exp(12.174) = 1,93,687.2 \\ X = 51; Y &= \exp(11.804 + 0.0074 \times 51) = \exp(12.1814) = 1,95,125.80\end{aligned}$$

The difference in the average cost of treatment for patients aged 50 and 51 is INR 1438.602 (note that the regression coefficient values are truncated after 3 decimals, inclusion of more decimals will give slightly different answer).

Alternatively,  $e^{11.804+0.0074X} (e^{0.0074} - 1) = 1438.602$

- Using the model chosen in question 1, is it possible to conclude that a patient weighing 51 kg is likely to spend at least INR 500 more than the one weighing 50 kg at 10% significance ( $\alpha = 0.10$ )?

**Answer:** To answer this question, we have to conduct one-tailed *t*-test. Before that, we have to calculate, for what minimum value of  $\alpha_1$ , the difference in the cost of treatment for patients aged 51 and 50 will be INR 500.

From the model:  $\alpha_0 = 11.804$  and  $\alpha_1 = 0.0074$ .

$$\ln(Y|X = 50) = 11.804 + 0.0074 \times 50 = 12.1740$$

Thus, the treatment cost of patient aged 50 is

$$\exp(12.1740) = Y = 1,93,687.2454$$

Assume that the treatment cost for patient aged 51 is at least 500 more than the patient aged 50, that is, for  $Y$  is at least 1,94,187.2454 when  $\alpha_1 = 0.0074$ . We have to find  $\alpha$  such that  $\exp(11.804 + 51 \alpha) - \exp(11.804 + 50 \alpha)$  is at least 500. We can use Microsoft Excel Solver to find the value of alpha that satisfies the above condition and the corresponding value is 0.003182 (whereas the estimated value is 0.0074).

**Continued...**

The null and alternative hypotheses for the  $t$ -tests are given by

$$H_0: \alpha_1 \leq 0.003182$$

$$H_A: \alpha_1 > 0.003182$$

If the null hypothesis is true, then the difference between the average treatments costs for someone aged 51 and 50 will be less than INR 500. The corresponding test statistic is as follows:

$$t = \frac{0.0074 - 0.003182}{0.003} = 1.406$$

The  $t$ -critical value for  $\alpha = 0.1$  and  $df = 118$  is 1.2888. Since the  $t$ -statistics is greater than  $t$ -critical, we reject the null hypothesis. The  $p$ -value (in a one-tailed  $t$ -test) corresponding to the  $t$ -value of 1.406 is 0.0811. Since the  $p$ -value is less than 0.1 (10% significance), we will reject the null hypothesis. Thus we conclude that the value of  $\alpha_1$  is greater than 0.003182 and so the difference between the cost of treatment for patients aged 51 and 50 is at least 500 at 10% significance (or 90% confidence).

5. At the time of admission, a patient's body weight is 50 kg. At 95% confidence level ( $\alpha = 0.05$ ), what will be the maximum cost of treatment for this patient?

**Answer:** The 95% confidence interval for  $\ln(Y_i)$  is given by

$$\ln(\hat{Y}_i) \pm t_{\alpha/2, n-2} \times S_e \times \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

$$\ln(\hat{Y}_i) = 11.8040 + 0.0074 \times 50 = 12.174 ; t_{0.025, 118} = 1.9802, S_e = 0.3975$$

$$\bar{X} = 56.2333, (X_i - \bar{X})^2 = (50 - 56.2333)^2 = 38.8544, \sum_{i=1}^{120} (X_i - \bar{X})^2 = 15979.4667$$

Substituting the aforementioned values in the equation, we get

$$12.174 \pm 1.9802 \times 0.3975 \times \sqrt{1 + \frac{1}{120} + \frac{38.8544}{15979.4667}} = (11.3826, 12.9653)$$

Note that the aforementioned prediction interval is for the  $\ln(Y)$ . The prediction interval for  $Y$  will be  $[\exp(11.3826), \exp(12.9653)]$ , that is the prediction interval for the treatment cost  $Y$  is (87780.97, 427324.9). So the maximum cost of treatment at 95% is  $\exp(12.9653) = 427324.9$ .

6. For a patient weighing 50 kg, what is the 95% confidence interval for the average cost of treatment?

**Answer:** The 95% confidence interval for the average value of the response variable is given by

$$\ln(\hat{Y}_i) \pm t_{\alpha/2, n-2} \times S_e \times \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

**Continued...**

$$\ln(\hat{Y}_i) = 11.8040 + 0.0074 \times 50 = 12.174 ; t_{\alpha/2, 118} = 1.9802, S_e = 0.3975$$

$$\bar{X} = 56.2333, (X_i - \bar{X})^2 = (50 - 56.2333)^2 = 38.8544, \sum_{i=1}^{120} (X_i - \bar{X})^2 = 15979.4667$$

Substituting the aforementioned values in the equation, we get

$$12.174 \pm 1.9802 \times 0.3975 \times \sqrt{\frac{1}{120} + \frac{38.8544}{15979.4667}} = (12.0923, 12.2556)$$

Note that the aforementioned prediction interval is for the  $\ln(Y)$ , so the 95% confidence interval for the average cost of treatment for a patient aged 50 years is  $[\exp(12.0923), \exp(12.2556)] = [178497.4, 210169.7]$ .

7. DAD hospital is planning to introduce package price (flat fee) for the treatment and they would like to charge INR 3,00,000 for the patients weighing 50 kg. What is the probability that the treatment cost is likely to exceed the package price?

**Answer:** Note that in the population,  $\ln(Y)$  follows a normal distribution with mean  $11.8040 + 0.0074 \times 50 (= 12.174)$  and the standard deviation 0.3975. If the hospital is planning to charge INR 3,00,000 for patients aged 50 years, we have to find the probability that the actual cost of treatment would be higher than INR 3,00,000.

Note that  $\ln(3,00,000) = 12.61154$ . So the probability that the cost of treatment will exceed INR 3,00,000 is given by  $P[\ln(Y) \geq 12.61154]$ . We know that the mean is 12.174 and the corresponding standard deviation is 0.3975. Using normal distribution, we can get the corresponding probability as 0.1355 [in Microsoft Excel, the probability =  $1 - \text{Normdist}(12.61154, 12.174, 0.3975, \text{TRUE})$ ]. That is, if DAD charges INR 3,00,000 for a patient aged 50 years, in 13.55% of the cases the actual cost of treatment is likely to exceed INR 3,00,000.

## SUMMARY

1. Regression is an important technique in predictive analytics. The primary objective of regression is to establish the existence of an association relationship between an outcome variable and predictor variables.
2. In a simple linear regression, the number of independent variable is one. It is used to understand the association relationship between KPIs and factors that may influence KPIs.
3. Linear regression implies that the relationship between the dependent variable and regression parameters  $\beta_0$  and  $\beta_1$  is linear. The relationship between the dependent variables and independent variable can be non-linear.
4. Regression parameters are estimated using the method of least squares under the assumption that the residuals follow a normal distribution. Method of least squares gives the best linear unbiased estimate.
5. Once a regression model is developed, it has to be validated using measures such as  $R^2$  and hypothesis tests.  $R^2$  is the percentage of variation in dependent variable explained by the model, whereas t-test is used for checking statistical significance of the independent variable.
6. Residual analysis is performed to check whether the model satisfies assumptions such as normality of residuals and homoscedasticity.
7. Once the model is accepted after validation tests, actionable items are derived from the model for model deployment.

### MULTIPLE CHOICE QUESTIONS

1. Regression models cannot be used for
  - (a) Analysing time-series data
  - (b) Understanding association relationship
  - (c) Understanding cause and effect relationship
  - (d) All of the above
2. The best simple linear regression model is the one for which
  - (a) The  $R$ -square (coefficient) is the highest.
  - (b) The residuals follow normal distribution.
  - (c) The  $p$ -value corresponding to  $t$ -test is less than the significance value  $\alpha$ .
  - (d) The  $p$ -value corresponding to  $t$ -test is less than the significance value  $\alpha$  and the residuals follow normal distribution and the residual are homoscedastic.
3. Which of the following equations are linear regression models?
  - (a)  $Y = \beta_0 + \beta_1 X^2$
  - (b)  $Y = \beta_0 + [1/(1+\beta_1)] X$
  - (c)  $Y = \beta_0 + \beta_1 X$
  - (d)  $\ln(Y) = \beta_0 + \beta_1 \ln(X)$
4. A high street jewellery shop uses a regression model  $Y = -10.5 + 95 \times \text{carat}$  to predict the price of a diamond as a function of carat, where carat is the weight of the diamond. The value of  $\beta_0$  is negative because:
  - (a) Regression model is incorrect since the value of diamond cannot take negative value.
  - (b) The regression models cannot be extrapolated beyond the range of the data used for building the model.
  - (c) The regression model is valid only for carat values greater than 0.1106 since the value of  $Y$  will be positive when carat is greater than 0.1106.
  - (d) The value of  $\beta_0$  ( $= -10.5$ ) should be ignored while calculating the price of the diamond.
5. If the residuals do not follow normal distribution:
  - (a) The regression coefficient estimates are incorrect.
  - (b) The  $R$ -square values are incorrect.
  - (c) The standard error of estimate is incorrect.
  - (d) The  $t$ -test for the coefficient of the explanatory variable ( $\beta_i$ ) is not valid.
6. If the correlation between a predictor variable and the outcome variable is 0.8, the proportion of variation in the outcome variable explained by the predictor variable is
  - (a) 0.9
  - (b) 0.72
  - (c) 0.89
  - (d) 0.64
7. Heteroscedasticity of the residual implies
  - (a) The variance of error for different values of the explanatory variables is different.
  - (b) The variance of error for different values of the explanatory variables is same.
  - (c) The variance of error decreases and the value of explanatory variable increases.
  - (d) The variance of error increases as the value of outcome variable increases.
8. In a model  $\ln(Y) = \beta_0 + \beta_1 X$ , the value of  $\beta_1$  is
  - (a) Change in value of  $Y$  for unit change in value of  $X$ .
  - (b) Change in value of  $X$  for unit change in value of  $Y$ .
  - (c) Percentage change in value of  $X$  for unit change in value of  $Y$ .
  - (d) Percentage change in value of  $Y$  for unit change in value of  $X$ .
9. Mahalanobis distance is a
  - (a) Measure of performance of the regression model.
  - (b) Measure of outlier.
  - (c) Measure of error.
  - (d) Measure of explained variation.
10. Transformation of outcome variable and predictor variable is used for
  - (a) Improving coefficient of determination.
  - (b) Removing heteroscedasticity
  - (c) Removing patterns in residual plot
  - (d) All of the above

**EXERCISES**

1. For a simple linear regression, prove the following relationship between  $F$ -statistic and  $R^2$ :

$$F = \frac{(n - 2) \times R^2}{(1 - R^2)}$$

2. In a simple linear regression model, prove that the value of  $F$ -statistic is same as the square of  $t$ -statistic value (that is,  $F = t^2$ ).
3. Price of a diamond is determined by 4Cs, namely, Carat, Cut, Clarity and Colour. Carat is the weight of the diamond, and 1 carat is equivalent to 0.2 grams. Data on carat and price of 6000 diamonds are used for developing SLR models. The mean and the standard deviation of diamond price and carat are provided in Table 9.15.

**TABLE 9.15** Descriptive statistics

	Carat	Price
Mean	1.33	11792
Standard deviation	0.48	10184

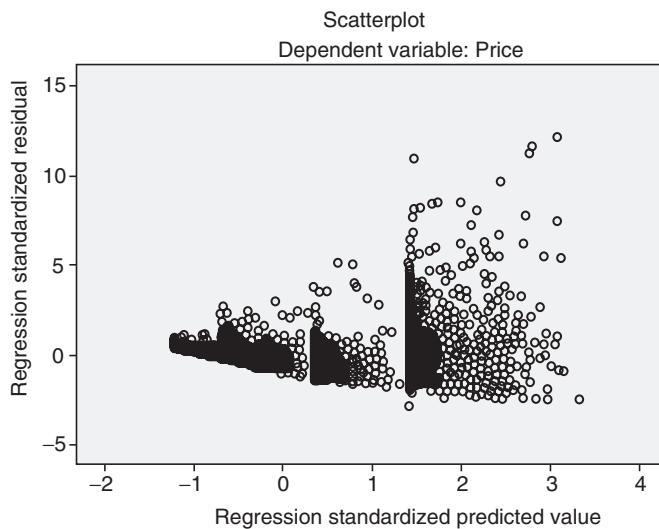
A regression model (model 1) based on data of 6000 diamonds is developed using price as the dependent variable and carat as the independent variable.

$$\text{Model 1: } Y = \beta_0 + \beta_1 \times \text{Carat}$$

The SPSS output for model 1 and the corresponding residual plot is shown in Table 9.16 and Figure 9.15, respectively.

**TABLE 9.16** Regression co-efficient

Model	Unstandardized Coefficients		Standardized Coefficients		<i>t</i> -value	Sig.
	<i>B</i>	Std. Error	Beta			
1	(Constant)	-12738.581	200.801		-63.439	.000
	Carat	18381.261	141.733			

**FIGURE 9.14** Plot between standardized predicted value versus standardized residual for model 1.

- (a) Based on the values in Tables 9.15 and 9.16 and Figure 9.14, comment on the validity of the model 1.
- (b) Is it possible to conclude that the price of the diamond increases by at least 10,000 for every one-carat increase in the diamond weight (at significance value of 0.05)? State clearly the condition(s) under which the above claim is true.

$$\text{Model 2: } \ln(Y) = \beta_0 + \beta_1 \times \text{Carat}$$

The SPSS regression output for model 2 is shown in Tables 9.17 and 9.18. The residual plot is shown in Figure 9.15.

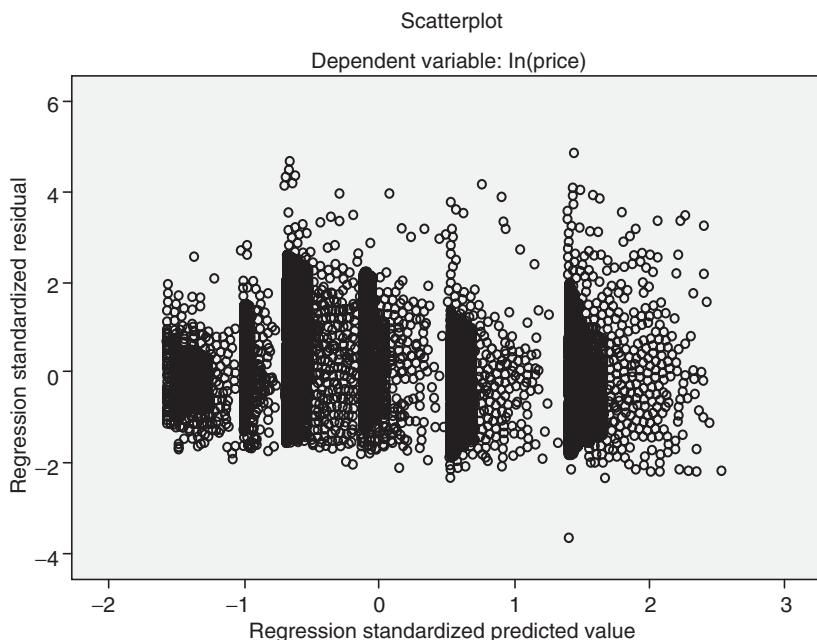
**TABLE 9.17** Model summary

Model	R	R-Square	Standard Error of Estimate
1	0.921	0.848	0.2768907

**TABLE 9.18** Coefficients<sup>a</sup>

Model	Unstandardized Coefficients		Standardized Coefficients		t	Sig.
	B	Std. Error	Beta			
1	(Constant)	7.265	0.011		682.302	0.000
	Carat	1.375	0.008	0.921	183.004	0.000

<sup>a</sup>Dependent variable: ln(Price).



**FIGURE 9.15** Plot between standardized value versus standardized residual for model 2.

- (c) What is the interpretation of the coefficient for the variable carat in model 2?
- (d) Calculate the maximum possible price of a specific diamond whose weight is 0.4 grams at 95% confidence level using model 2.
- (e) Between models 1 and 2, which model should be used to explain variation in the diamond price? State the reasons clearly.

4. Table 9.19 provides the winning margin of all 20 Lok Sabha constituencies of Kerala in 2014 parliament elections of India and maximum delay of top 20 flights (origin–destination) of Air India between 15 July 2014 and 15 September 2014.

**TABLE 9.19** Data on Lok Sabha election winning margin of Kerala constituencies and maximum delay of top 20 Air India flights

S. No.	Constituency	Winning Margin	Air India Top 20 flights	Maximum Delay in Minutes
1	Alappuzha	19407	Bangalore–Mumbai	182
2	Alathur	37312	Ahmedabad–Mumbai	203
3	Attingal	69378	Hyderabad–Mumbai	240
4	Chalakudy	13884	Mumbai–Goa	164
5	Ernakulum	87047	Delhi–Kolkata	265
6	Idukki	50542	Chennai–Delhi	226
7	Kannur	6566	Delhi–Bangalore	156
8	Kasaragod	6921	Mumbai–Chennai	161
9	Kollam	37649	Kolkata–Delhi	219
10	Kottayam	120599	Mumbai–Delhi	328
11	Kozhikode	16883	Hyderabad–Delhi	181
12	Malappuram	194740	Delhi–Mumbai	340
13	Mavelikkara	32737	Mumbai–Ahmedabad	202
14	Palakkad	105300	Mumbai–Hyderabad	284
15	Pathanamthitta	56191	Chennai–Mumbai	234
16	Ponnani	25410	Bangalore–Delhi	199
17	Thiruvananthapuram	15470	Goa–Mumbai	178
18	Thrissur	38228	Delhi–Chennai	225
19	Vadakara	3306	Delhi–Hyderabad	146
20	Wayanad	20870	Mumbai–Bangalore	197

Data Source: [www.flightstatus.com](http://www.flightstatus.com)

- (a) Develop a simple linear regression model between winning margin ( $Y$ ) and maximum flight delay ( $X$ ) and calculate the regression coefficients.  
 (b) What is the value of  $R^2$ ?  
 (c) Is the model statistically significant, what can you infer from the regression model?
5. The box-office collection of a Bollywood movie across different regions and the corresponding social media engagement (likes + dislikes) is provided in Table 9.20.

**TABLE 9.20** Social media engagement versus box-office collection

Region	Cumulative Likes + Dislikes (Engagement)	Revenue (INR)
Mumbai Territory	908104	70,056,138
Delhi/UP	1885487	45,230,603
East Punjab	845910	17,193,472

**TABLE 9.20** Social media engagement versus box-office collection—Continued

Region	Cumulative Likes + Dislikes (Engagement)	Revenue (INR)
West Bengal	1071577	15,074,364
Bihar	5	6,165,934
Rajasthan	3188	11,934,830
Nizam/AP	11527	14,984,099
Mysore	189588	5,923,729
Assam	34939	2,371,340
Odisha	999024	2,328,932
TNK	644074	1482738
CP	482457	14,224,686
CI	296348	10,595,171

- (a) Develop a simple linear regression model for the data shown in Table 9.20. Is there any evidence that the box-office collection ( $Y$ ) of the movie has statistically significant relationship with the social media engagement ( $X$ )?
- (b) What is the 95% confidence interval for the average box-office collection for a movie with 20,000 likes and dislikes?
- (c) Should Bollywood movie producers invest more to promote their movies through social media?
6. Corruption perception index (source: Transparency International) and Gini Index (Source: Wikipedia) of 20 countries is shown in Table 9.21. Corruption perception index close to 100 indicates low corruption and close to 0 indicates high corruption. Gini index is a measure of income distribution among citizens of a country (high Gini indicates high inequality).

**TABLE 9.21** Corruption Index and Gini Index

Country	Corruption Index	Gini Index
Hong Kong	77	53.7
South Korea	53	30.2
China	40	46.2
Italy	47	32.7
Mongolia	38	36.5
Austria	75	27.6
Norway	85	23.5
UK	81	31.6
Canada	82	33.7
Germany	81	30.7
Sweden	88	25.4
Denmark	90	27.5
France	69	30.1

**TABLE 9.21** Corruption Index and Gini Index—Continued

Country	Corruption Index	Gini Index
United States	74	40.8
Russia	29	40.1
Portugal	62	34.2
Romania	48	34
Argentina	36	42.7
Greece	44	34.2
Thailand	35	39.4

- (a) Develop a simple linear regression model ( $Y = \beta_0 + \beta_1 X$ ) between corruption perception index ( $Y$ ) and Gini index ( $X$ ). What is the change in the corruption perception index for every one unit increase in Gini index?
- (b) What proportion of the variation in corruption perception index is explained by Gini index?
- (c) Is there a statistically significant relationship between corruption perception index and Gini index at  $\alpha = 0.1$ ?
- (d) Calculate the 95% confidence interval for the regression coefficient  $\beta_1$ .
- (e) Is it possible to conclude that the corruption perception index will decrease by at least 1 unit for every one unit increase in Gini index? Conduct an appropriate hypothesis test at  $\alpha = 0.05$ .
- (f) Calculate 95% confidence interval for the expected value of corruption perception index for Gini index value = 30.
7. A regression model is developed between corruption perception index and per capita income (in US dollars) based on data on 20 countries. Regression model output obtained through Microsoft Excel is shown in Table 9.22. Note that Table 9.22 shows only partial output of the model developed.

**TABLE 9.22** Regression between corruption perception index ( $Y$ ) and per capita ( $X$ )**SUMMARY OUTPUT**

Regression Statistics	
Multiple R	
R Square	
Adjusted R Square	
Standard Error	10.94929
Observations	20

**ANOVA**

	df	SS	MS	F	Significance F
Regression	1	5918.236			
Residual	18	2157.964			
Total					
	Coefficients	Standard Error	t-Stat	p-value	Lower 95%
Intercept	6.496415				5.773095
Per Capita	0.00016				0.000788
					0.001461

- (a) What proportion of the corruption perception index is explained by per capita?
- (b) What is change in the value of corruption perception index for every one dollar increase in per capita?
- (c) Is there a statistically significant relationship between corruption perception index and per capita at  $\alpha = 0.01$ ?
- (d) What is the average corruption perception index when per capita is \$ 30,000. What is the corresponding 95% confidence interval?
- (e) Per capita of a country is \$ 30,000. What is the probability that the corruption perception index of this country is less than 50?
- (f) Which of the following statements are true based on the model shown in Table 9.21?
  - (i) Corruption perception index and per capita are positively correlated.
  - (ii) Corruption perception index and per capita are negatively correlated.
  - (iii) There is no correlation between corruption perception index and per capita.

---

## REFERENCES

1. Hanley J A, (2004), "Transmuting women into men: Galton's Family Data on Human Stature", *The American Statistician*, **58**(3), 237–243.
2. Harter H L, (1974), "Method of Least Squares and Some Alternatives – Part II", *International Statistical Review*, **42**(3), 235–282.
3. Harter H L, (1975), "Method of Least Squares and Some Alternatives – Part IV", *International Statistical Review*, **43**(2), 125–190.
4. Hodge G (2012), "The Ugly Truth of Online Dating: Top 10 Lies Told by Internet Daters", *The Huffington Post*, 10 October 2012, available at [http://www.huffingtonpost.com/greg-hodge/online-dating-lies\\_b\\_1930053.html](http://www.huffingtonpost.com/greg-hodge/online-dating-lies_b_1930053.html), accessed on 21 April 2017.
5. Galton F (1886), "Regression Towards Mediocrity in Hereditary Stature", *The Journal of the Anthropological Institute Great Britain and Ireland*, **15**, 246–263.
6. Kutner M H, Nachtsheim C J, Neter J and Li W (2013), "Applied Linear Models", Fifth Edition, McGraw Hill, New Delhi.
7. Ryan T P (2009), "Modern Regression Methods – 2nd Edition", John Wiley and Sons, Hoboken, New Jersey.
8. Waugh F V (1961), "The Place of Least Squares in Econometrics", *Econometrica*, **29**(3), 386–396.



# Multiple Linear Regression

10

“Regression is like a woman. You never understand it fully. When you think you do, the next model is entirely different”.

— Daksh Itha Wickramasinghe

## LEARNING OBJECTIVES

- LO 10-1** Understand the difference between simple linear regression and multiple linear regression (MLR).
- LO 10-2** Understand various stages in MLR model building and the underlying assumptions of multiple regression models.
- LO 10-3** Learn to incorporate categorical (qualitative) variables in MLR and the use of dummy variables in MLR model building.
- LO 10-4** Interpretation of multiple regression parameters and the concept of partial regression coefficients.
- LO 10-5** Understand the impact of multi-collinearity, auto-correlation on regression model.
- LO 10-6** Learn to build regression models using forward, backward, and stepwise regression.
- LO 10-7** Application of multiple regression models across several industries.

## ESSENCE OF MULTIPLE LINEAR REGRESSION

Multiple linear regression (MLR) is a statistical technique for finding existence of an association relationship between a dependent variable (aka response variable or outcome variable) and several independent variables (aka explanatory variables or predictor variable).

## 10.1 | INTRODUCTION

Key performance indicators (KPIs) of organizations may be simultaneously associated with many factors. For example, revenue generated from a product sold by a company may be simultaneously associated with factors such as price, market size, promotions, competitor's price, and so on. In such cases, we try to establish the existence of a relationship between a dependent variable and several independent variables. Multiple Linear Regression (MLR) model is a statistical model that establishes existence of a linear relationship (association) between a dependent variable and several independent variables.

The functional form of MLR is given by

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i \quad (10.1)$$

In Eq. (10.1) the variable  $Y$  is the dependent variable (response variable or outcome variable);  $X_1, X_2, \dots, X_k$  are independent variables (predictor variables or explanatory variables);  $\beta_0$  is a constant;  $\beta_1, \beta_2, \dots, \beta_k$  are called the partial regression co-efficients corresponding to the explanatory variables  $X_1, X_2, \dots, X_k$ , respectively; and  $\varepsilon_i$  is the error term (or residual). The model in Eq. (10.1) is called a **response surface (hyperplane)** that can be complex depending on the functional form of the relationship. In the matrix form, Eq. (10.1) can be written as

$$Y = X\beta + \varepsilon \quad (10.2)$$

where  $Y$  is a vector of response variables of size  $(N \times 1)$ ,  $X$  is a matrix of explanatory variable values with size  $(N \times k + 1)$ ,  $\beta$  is a vector of regression parameters of size  $(k + 1 \times 1)$ ,  $\varepsilon$  is a vector of errors that follows normal distribution  $N(0, \sigma^2)$  and the size is  $(N \times 1)$ , where  $N$  is the sample size and  $k$  is the number of independent variables in the model. Elements of matrix representation of the multiple regression are provided below:

$$\begin{aligned} Y &= \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} & X &= \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & \cdots & X_{1,k} \\ 1 & X_{2,1} & X_{2,2} & & & X_{2,k} \\ \vdots & \vdots & \vdots & & & \vdots \\ 1 & X_{N,1} & X_{N,2} & & & X_{N,k} \end{bmatrix} \\ \beta &= \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} & \varepsilon &= \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix} \end{aligned}$$

**IMPORTANT**

The regression coefficients  $\beta_1, \beta_2, \dots, \beta_k$  are called partial regression coefficients since the relationship between an explanatory variable and the response variable is calculated after removing (partial out) the effect of all the other explanatory variables in the model.

## 10.2 | ORDINARY LEAST SQUARES ESTIMATION FOR MULTIPLE LINEAR REGRESSION

Consider the MLR model with  $n$  independent variables as given by Eq. (10.1). A few of these explanatory variables may be derived variables (such as  $X_i^2, X_i X_j$ , and so on). The assumptions of multiple linear regression model are as follows:

1. The regression model is linear in parameter.
2. The explanatory variable,  $X_i$ , is assumed to be non-stochastic (that is,  $X_i$  is deterministic).

3. The conditional expected value of the residuals,  $E(\varepsilon_i|X_i)$ , is zero.
4. In a time series data, residuals are uncorrelated, that is,  $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$  for all  $i \neq j$ .
5. The residuals,  $\varepsilon_i$ , follow a normal distribution.
6. The variance of the residuals,  $\text{Var}(\varepsilon_i|X_i)$ , is constant for all values of  $X_i$ . When the variance of the residuals is constant for different values of  $X_i$ , it is called **homoscedasticity**. A non-constant variance of residuals is called **heteroscedasticity**.
7. There is no high correlation between independent variables in the model (called **multicollinearity**). Multi-collinearity can destabilize the model and can result in incorrect estimation of the regression parameters.

The sum of squared errors for the model in Eq. (10.1) is given by

$$SSE = \sum_{i=1}^N \varepsilon_i^2 = \sum_{i=1}^N (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^N [Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki})]^2 \quad (10.3)$$

Differentiating Eq. (10.3) with respect to various regression coefficients in the model, we get

$$\frac{\partial SSE}{\partial \beta_0} = -2 \sum_{i=1}^N [Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki})] \quad (10.4)$$

$$\frac{\partial SSE}{\partial \beta_1} = -2 \sum_{i=1}^N [Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki})] \times X_{1i} \quad (10.5)$$

.... .... .... .... .... .... ....

$$\frac{\partial SSE}{\partial \beta_k} = -2 \sum_{i=1}^N [Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki})] \times X_{ki} \quad (10.6)$$

To find the minimum SSE defined in Eq. (10.3), we set the system of equations defined in Eqs. (10.4) to (10.6) to zero:

$$\frac{\partial SSE}{\partial \beta_0} = -2 \sum_{i=1}^N [Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki})] = 0 \quad (10.7)$$

$$\frac{\partial SSE}{\partial \beta_1} = -2 \sum_{i=1}^N [Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki})] \times X_{1i} = 0 \quad (10.8)$$

.... .... .... .... .... .... ....

$$\frac{\partial SSE}{\partial \beta_k} = -2 \sum_{i=1}^N [Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki})] \times X_{ki} = 0 \quad (10.9)$$

Simplifying the first-order conditions (equating 1<sup>st</sup> derivative to zero to find the minimum value of SSE) in Eqs. (10.7)–(10.9), we get the following linear system of equations:

$$N\beta_0 + \beta_1 \sum_{i=1}^N X_{1i} + \beta_2 \sum_{i=1}^N X_{2i} + \dots + \beta_k \sum_{i=1}^N X_{ki} = \sum_{i=1}^N Y_i \quad (10.10)$$

$$\beta_0 \sum_{i=1}^N X_{1i} + \beta_1 \sum_{i=1}^N X_{1i}^2 + \beta_2 \sum_{i=1}^N X_{2i} X_{1i} + \dots + \beta_k \sum_{i=1}^N X_{ki} X_{1i} = \sum_{i=1}^N Y_i X_{1i} \quad (10.11)$$

.....   .....

$$\beta_0 \sum_{i=1}^N X_{ki} + \beta_1 \sum_{i=1}^N X_{1i} X_{ki} + \beta_2 \sum_{i=1}^N X_{2i} X_{ki} + \dots + \beta_k \sum_{i=1}^N X_{ki}^2 = \sum_{i=1}^N Y_i X_{ki} \quad (10.12)$$

Solution to the system of linear equations in Eqs. (10.10)–(10.12) will give explicit expressions for the **partial regression coefficients**  $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ . The system of equations can be written in the matrix form as follows:

$$\begin{bmatrix} N & \sum_{i=1}^N X_{1i} & \dots & \sum_{i=1}^N X_{ki} \\ \sum_{i=1}^N X_{1i} & \sum_{i=1}^N X_{1i}^2 & \sum_{i=1}^N X_{1i} X_{ki} & \\ \dots & \dots & \dots & \dots \\ \sum_{i=1}^N X_{ki} & \sum_{i=1}^N X_{1i} X_{ki} & \sum_{i=1}^N X_{ki}^2 \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^N Y_i \\ \sum_{i=1}^N X_{1i} Y_i \\ \vdots \\ \sum_{i=1}^N X_{ki} Y_i \end{bmatrix} \quad (10.13)$$

Solving the system of equations described in Eq. (10.13), we get the estimated values of the regression coefficients,  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ . Note that matrix  $X$  is not a square matrix and to solve of the regression coefficient we have to make the matrix a square matrix by multiplying it with  $X^T$  (transpose of  $X$ ); thus, Eq. (10.2) can be written as

$$X^T Y = X^T X \hat{\beta} \quad (10.14)$$

The regression coefficients  $\hat{\beta}$  is given by

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (10.15)$$

The estimated values of response variable are [combining Eqs. (10.2) and (10.15)]

$$\hat{Y} = X \hat{\beta} = X(X^T X)^{-1} X^T Y \quad (10.16)$$

Note that in Eq. (10.16) the predicted value of dependent variable  $\hat{Y}_i$  is a linear function of  $Y_i$ . Equation (10.16) can be written as follows:

$$\hat{Y} = HY \quad (10.17)$$

$H = X(X^T X)^{-1} X^T$  is called the **hat matrix**, also known as the **influence matrix**, since it describes the influence of each observation on the predicted values of the response variable. Hat matrix plays a crucial role in identifying the outliers and influential observations in the sample.

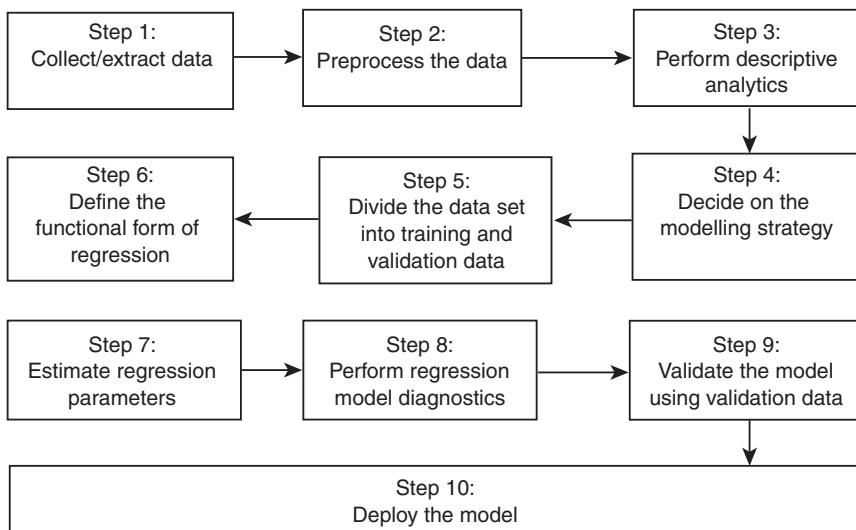
### 10.3 | MULTIPLE LINEAR REGRESSION MODEL BUILDING

A few examples of MLR are as follows:

1. The treatment cost of a cardiac patient may depend on factors such as age, past medical history, body weight, blood pressure, and so on.
2. Salary of MBA students at the time of graduation may depend on factors such as their academic performance, prior work experience, communication skills, and so on.
3. Market share of a brand may depend on factors such as price, promotion expenses, competitors' price, etc.

In many cases, the potential number of independent variables can run into thousands, especially in industry sectors such as healthcare, banking and finance, and telecom. Although the data sets may have several variables, managers may like to identify few variables (preferably in single digits) that can be used for prediction. This means identifying the most important predictor variables among hundreds of potential predictor variables. Knowing the top few predictors helps the decision makers to focus on them to manage the KPIs. The framework for multiple regression model building is described in Figure 10.1.

The framework for developing a multiple linear regression model is similar to that of a simple linear regression model; however, the data scientist has to handle issues such as multicollinearity, over-fitting, etc. The activities performed during various stages of MLR model development are discussed next section. MLR model development is an iterative process and many steps in Figure 10.1 may be repeated to find the best MLR model while dealing with real-life problems.



**FIGURE 10.1** Framework for building multiple linear regression (MLR).

**STEP 1** *Collect/Extract Data*

The first step in regression model is to collect and/or extract data for the problem identified. In case of multiple linear regression model, we will have several independent variables. It is not uncommon to have more than 1000 independent variables for a problem in real-life cases. Many of these variables are derived variables (will be explained in the next paragraph).

**STEP 2** *Pre-process the Data*

The issues that are handled during the pre-processing stage are as follows:

1. **Data Quality (measured through several characteristics such as completeness, correctness, etc.):** Data completeness refers to availability of necessary data for developing the model. Consider a manufacturing company that would like to reduce the warranty costs on one of their products. The company may have captured the number of failures (thus warranty claims) of their products during the warranty period. If the objective is to reduce the warranty costs, then the data related to time between failures of each component of the product should be captured so that the time to failure distribution and causes of failures can be identified which in turn can result in actionable items in the form of design changes. If the data set has only the number of claims and not the actual failure times of all components and the failure causes, we may not be able to come up with effective actions to reduce the warranty costs.
2. **Missing Data:** Many variables may have missing values. The data scientist has to come up with a strategy to handle missing values such as data imputation and specific techniques to carry out the imputation.
3. **Handling Qualitative Variables:** Qualitative variables or categorical variables need to be converted using dummy variables before incorporating them in regression model.
4. **Derive new variables** (such as ratios and interaction variables), which may have better association relationship with the dependent variable.

In multiple regression models, many derived variables and transformations can be used to explain the response variable. For example, assume that the decision maker is developing a credit risk model of housing loan applicants. Assume that the decision maker has collected factors such as loan amount (say,  $X_1$ ) and value of the property (say,  $X_2$ ). Instead of using  $X_1$  and  $X_2$  as two independent variables in the model, it may be better to use the ratio  $X_2/X_1$  (value/loan) as the independent variable. If the data set has 20 continuous independent variables, then there will be  ${}^{20}C_2$  ( $=190$ ) possible ratios. Since many analytics problems are likely to have several variables, the derived variables such as ratios can increase the total number of predictors significantly.

Another pre-processing that may be required is **interaction variables** (interaction variables are products of a two variables,  $X_1X_2$  – we will discuss this in detail later).

**STEP 3** *Perform Descriptive Analytics*

It is a good practice to start the MLR model building with descriptive analytics. In addition to descriptive statistics and data visualizations such as scatter plot and box plot, it is useful to check correlation between different variables since it can provide early warning for issues such as multi-collinearity.

It will be also useful to identify proxy variables in case an original variable cannot be used for some reason. For example, a vehicle insurance company found that among new car drivers, young male drivers had higher average claim compared to young female drivers. Insurance companies would like to charge higher premium for young male drivers since the risk of claim is higher from young male drivers compared to young female drivers. However, in many countries discrimination based on gender may not be permitted legally. So, one has to identify a proxy for gender. While analysing the data the company found that there was a clear difference between car models driven by male and female drivers. The insurance company can charge different premium for different car models, to differentiate premium charged for male and female drivers since discrimination based on car model is not under the purview of law. In this case, the insurance premium was thus fixed based on the car models used by the drivers which was used as a proxy variable for gender.

**STEP 4** *Modelling Strategy*

When the number of variables runs into several hundreds, building regression models can get complicated due to multi-collinearity as well as computational complexity since estimation of regression parameters involves matrix inversion (Hat Matrix). The data scientist may decide to use various data reduction techniques to reduce the number of variables, such as Principal Component Analysis. The data scientist may also use specific variable selection approaches such as **Forward Selection, Backward Elimination or Stepwise Regression** (discussed later in the chapter). Use of Mallows's  $C_p$  is another strategy that a data scientist can use.

**STEP 5** *Divide the Data into Training and Validation Data*

The next step is to divide the data into two or more subsets to build and validate the model for final model selection. The data scientist may decide to use cross validation in which several training and validation sets are created randomly for the final model selection.

**STEP 6** *Define the Functional Form*

Most data scientists may start with a linear relationship between the dependent and the independent variables. However, the functional form may be changed if there is a lack of fit.

**STEP 7** Estimate Regression Parameters

Once the functional form is specified, the next step is to estimate the partial regression coefficients using the method of **Ordinary Least Squares** (OLS). OLS is used to fit a polygon through a set of data points, such that the sum of the squared distances between the actual observations in the sample and the regression equation is minimized. OLS provides the **Best Linear Unbiased Estimate** (BLUE), that is,  $E[\beta - \hat{\beta}] = 0$ , where  $\beta$  is the population parameter and  $\hat{\beta}$  is the estimated parameter value from the sample.

**STEP 8** Perform Regression Model Diagnostics

Apart from checking the statistical significance of predictor variables, validating the multiple regression models also involves checking for issues such as multi-collinearity, heteroscedasticity, auto-correlation (in the case of time-series data), outlier analysis, etc. This is over and above checking for the normality of residuals.

*F*-test is used for checking the overall significance of the model whereas *t*-tests are used to check the significance of the individual variables. Presence of multi-collinearity can be checked through measures such as **Variance Inflation Factor** (VIF).

Remedial measures are necessary if any of the regression model assumptions are violated (the assumptions are discussed in Section 10.2). The remedial measures are discussed in Section 10.19.

**STEP 9** Validate the Model using Validation Data

Final model selection will depend on the performance of the model in validation data. The measures that can be used for validating the model in the validation data are as follows:

1.  $R^2$  or Adjusted  $R^2$
2. Mean absolute percentage error,  $\sum_{i=1}^K \frac{1}{K} \left| \frac{Y_i - \hat{Y}_i}{Y_i} \right| \times 100\%$ , where  $K$  is the number of cases in the validation data.
3. Root Mean Square Error (RMSE),  $\sqrt{\sum_{i=1}^K \frac{1}{K} (Y_i - \hat{Y}_i)^2}$ .

Models that give consistent performance in both training and validation data set will be chosen for deployment.

**STEP 10** Deploy the Model

The final step in the regression model is to generate actionable items and the implementation plan. For example, computer apps can be developed based on the regression model so that the organization can

use it in an efficient manner. Automation of actionable items is one of the important tasks in analytics model deployment.

Multiple linear regression model development is an iterative process; the modeller may have to go through many iterations of model development before choosing the correct model. Also, every analytics model has a life and its performance may deteriorate due to changes in the market/environment. So data scientists have to monitor the model performance and may also have to change the model in case the model performance deteriorates.

---

## 10.4 | PART (SEMI-PARTIAL) CORRELATION AND REGRESSION MODEL BUILDING

The concept of semi-partial correlation (or part correlation) plays an important role in MLR model building. The increase in the coefficient of determination,  $R^2$ , when a new variable is added is given by the square of the semi-partial correlation of the newly added variable with dependent variable  $Y$ . In this section, we will be discussing the concept of partial and semi-partial correlations.

Consider a regression model with two independent variables (say  $X_1$  and  $X_2$ ). The model can be written as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \quad (10.18)$$

Here, the regression coefficients are partial regression coefficients, since to interpret  $\beta_1$  (or  $\beta_2$ ), we have to control the value of  $X_2$  (or  $X_1$ ). That is,  $\beta_1$  is the change in the value of  $Y$  when  $X_1$  is changed by one unit and  $X_2$  is kept constant. Assume that  $Y$  is the revenue generated from a product at a retail store,  $X_1$  is the price of the product, and  $X_2$  is the amount of snowfall in the neighbourhood where the store is located. In this case, when  $X_1$  changes by one unit,  $\beta_1$  is the change in the revenue provided the value of the snowfall is kept constant (or for a fixed snowfall value). Alternatively,  $\beta_1$  is the change in the value of revenue when the price of  $X_1$  is changed by one unit when the value of snowfall ( $X_2$ ) is controlled (which is probably difficult unless the store is owned by God).

When we develop the regression model, we try to capture the unique contribution of a variable in explaining the variation in the response variable. For better understanding of the multiple regression model building, we have to understand the partial and semi-partial correlations.

### 10.4.1 | Partial Correlation

Consider an MLR model between a response variable  $Y$  and two independent variables,  $X_1$  and  $X_2$ . Partial correlation is the correlation between the response variable  $Y$  and the explanatory variable  $X_1$  when influence of  $X_2$  is removed from both  $Y$  and  $X_1$  (in other words, when  $X_2$  is kept constant). Alternatively, partial correlation is the correlation between residualized response variable and residualized explanatory variables.

Let  $r_{YX_1, X_2}$  denote the partial correlation between  $Y$  and  $X_1$  when  $X_2$  is kept constant. Then  $r_{YX_1, X_2}$  is given by

$$r_{YX_1, X_2} = \frac{r_{YX_1} - r_{YX_2} \times r_{X_1 X_2}}{\sqrt{(1 - r_{YX_2}^2) \times (1 - r_{X_1 X_2}^2)}} \quad (10.19)$$

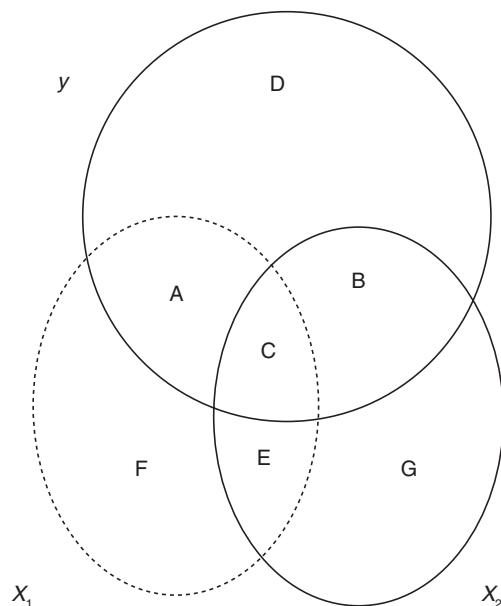
In Eq. (10.19),  $r_{YX_1}$  is the correlation coefficient (**Pearson Correlation**) between  $Y$  and  $X_1$ ,  $r_{YX_2}$  is the Pearson correlation between  $Y$  and  $X_2$ , and  $r_{X_1X_2}$  is the Pearson Correlation Coefficient between  $X_1$  and  $X_2$ . The Venn diagram in Figure 10.2 can be used as an analogy to understand the concept of partial correlation. The variation in variable  $Y$  can be divided into four components:  $A$ ,  $B$ ,  $C$ , and  $D$ . Removing the influence of  $X_2$  from both  $X_1$  and  $Y$  is equivalent to removing segments  $B$ ,  $C$ , and  $E$  from the Venn diagram. Once, we remove the influence of  $X_2$  from both  $X_1$  and  $Y$ , the segment that is common between  $Y$  and  $X_1$  is segment  $A$  (out of  $A + D$ ), that is, the ratio  $[A/(A + D)]$  is analogous to partial correlation.

#### 10.4.2 | Semi-Partial Correlation (or Part Correlation)

Consider a regression model between a response variable  $Y$  and two independent variables  $X_1$  and  $X_2$ . The semi-partial (or part correlation) between a response variable  $Y$  and independent variable  $X_1$  measures the relationship between  $Y$  and  $X_1$  when the influence of  $X_2$  is removed from only  $X_1$  but not from  $Y$ . It is equivalent to removing portions  $C$  and  $E$  from  $X_1$  in the Venn diagram shown in Figure 10.2 (but  $C$  will be retained in  $Y$ ). The ratio (or  $A/(A + B + C + D)$ ) is analogous the semi-partial correlation between  $Y$  and  $X_1$ .

Mathematically, semi-partial correlation between  $Y$  and  $X_1$ ,  $sr_{YX_1, X_2}$ , when influence of  $X_2$  is removed from  $X_1$  is given by

$$sr_{YX_1, X_2} = \frac{r_{YX_1} - r_{YX_2} r_{X_1X_2}}{\sqrt{(1 - r_{X_1X_2}^2)}} \quad (10.20)$$



**FIGURE 10.2** Partial and semi-partial correlation.

**IMPORTANT**

*Semi-partial (part) correlation plays an important role in regression model building. The increase in R-square (coefficient of determination), when a new variable is added into the model, is given by the square of the semi-partial correlation.*

**EXAMPLE 10.1**

The cumulative television rating points (*CTRP*) of a television program, money spent on promotion (denoted as *P*), and the advertisement revenue (in Indian rupees denoted as *R*) generated over one-month period for 38 different television programs is provided in Table 10.1. Develop a multiple linear regression model to understand the relationship between the advertisement revenue (*R*) generated as response variable and promotions (*P*) and *CTRP* as predictors.

**TABLE 10.1** Data on advertisement revenue (*R*) of programs along with *CTRP* and *P*

Serial	<i>CTRP</i>	<i>P</i>	<i>R</i>	Serial	<i>CTRP</i>	<i>P</i>	<i>R</i>
1	133	111600	1197576	20	156	104400	1326360
2	111	104400	1053648	21	119	136800	1162596
3	129	97200	1124172	22	125	115200	1195116
4	117	79200	987144	23	130	115200	1134768
5	130	126000	1283616	24	123	151200	1269024
6	154	108000	1295100	25	128	97200	1118688
7	149	147600	1407444	26	97	122400	904776
8	90	104400	922416	27	124	208800	1357644
9	118	169200	1272012	28	138	93600	1027308
10	131	75600	1064856	29	137	115200	1181976
11	141	133200	1269960	30	129	118800	1221636
12	119	133200	1064760	31	97	129600	1060452
13	115	176400	1207488	32	133	100800	1229028
14	102	180000	1186284	33	145	147600	1406196
15	129	133200	1231464	34	149	126000	1293936
16	144	147600	1296708	35	122	108000	1056384
17	153	122400	1320648	36	120	194400	1415316
18	96	158400	1102704	37	128	176400	1338060
19	104	165600	1184316	38	117	172800	1457400

The MLR model is given by

$$R \text{ (Advertisement Revenue)} = \beta_0 + \beta_1 \times CTRP + \beta_2 \times P \quad (10.21)$$

The regression coefficients can be estimated using OLS estimation. The SPSS output for the above regression model is provided in Tables 10.2 and 10.3.

**TABLE 10.2** Model summary

Model	R	R-Square	Adjusted R-Square	Std. Error of the Estimate
1	0.912 <sup>a</sup>	0.832	0.822	57548.382

<sup>a</sup>Predictors: (Constant),  $P$ ,  $CTRP$ .

**TABLE 10.3** Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1	Constant	41008.840	90958.920	0.451	0.655
	CTRP	5931.850	576.622	10.287	0.000
	P	3.136	0.303	10.344	0.000

The regression model after estimation of the parameters is given by

$$R = 41008.84 + 5931.850 \text{ } CTRP + 3.136P \quad (10.22)$$

For every one unit increase in  $CTRP$ , the revenue increases by 5931.850 when the variable promotion is kept constant, and for one unit increase in promotion the revenue increases by 3.136 when  $CTRP$  is kept constant. Note that television-rating point is likely to change when the amount spent on promotion is changed.

## 10.5 | INTERPRETATION OF MLR COEFFICIENTS – PARTIAL REGRESSION COEFFICIENT

In this section we will discuss how a multiple regression model is developed and the mathematics behind partial regression coefficient. Let us first start with the following simple linear regression model between revenue generated and  $CTRP$ :

$$R = \alpha_0 + \alpha_1 \times CTRP + \varepsilon_1 \quad (10.23)$$

Note that,  $\varepsilon_1$  is the variation in  $R$  (revenue generated through advertisement) not explained by  $CTRP$ . Tables 10.4 and 10.5 show simple linear regression SPSS output between  $R$  and  $CTRP$  [model defined in Eq. (10.23)].

**TABLE 10.4** Model summary

Model	R	R-Square	Adjusted R-Square	Std. Error of the Estimate
1	0.564	0.318	0.299	114293.708

**TABLE 10.5** Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients		t	Sig.
	B	Std. Error	Beta			
1	(Constant) 625763.106	141522.635			4.422	0.000
	CTRP 4569.214	1114.912	0.564		4.098	0.000

In the next model, we run a simple linear regression between promotions expenditure ( $P$ ) as dependent variable and  $CTRP$  as independent variable:

$$P = \delta_0 + \delta_1 \times CTRP + \varepsilon_2 \quad (10.24)$$

Here  $\varepsilon_2$  is the variation in  $P$  not explained by  $CTRP$ . The regression model output for model in Eq. (10.24) is shown in Tables 10.6 and 10.7.

**TABLE 10.6** Model summary

Model	R	R-Square	Adjusted R-Square	Std. Error of the Estimate
1	0.228	0.052	0.026	31635.39950

**TABLE 10.7** Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients		t	Sig.
	B	Std. Error	Beta			
1	(Constant) 186456.659	39172.105			4.760	0.000
	CTRP -434.495	308.597	-0.228		-1.408	0.168

The third model is between  $\varepsilon_1$  (variation in advertisement revenue not explained by  $CTRP$ ) and  $\varepsilon_2$  (variation in promotion expenditure not explained by  $CTRP$ ) as expressed in Eq. (10.25). The regression model SPSS output is provided in Tables 10.8 and 10.9:

**TABLE 10.8** Model summary

Model	R	R-Square	Adjusted R-Square	Std. Error of the Estimate
1	0.868	0.754	0.747	56743.46998

**TABLE 10.9** Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients		t	Sig.
	B	Std. Error	Beta			
1	(Constant)	-1.614E-010	9205.006		0.000	1.000
	Unstandardized Residual	3.136	0.299	0.868	10.491	0.000

$$\varepsilon_1 = \eta_0 + \eta_1 \times \varepsilon_2 + \varepsilon_3 \quad (10.25)$$

Note that, in Table 10.9, the regression coefficient  $\eta_1$  is 3.136, which is same as the partial regression coefficient value  $\beta_2$  in Eq. (10.22). That is, the regression coefficient  $\beta_2$  is obtained using a regression model in which the dependent variable is residualized revenue and the independent variable is residualized promotion expenses. Table 10.10 provides summary of all 3 models discussed in this section.

**TABLE 10.10** Model development in multiple linear regression

Model	Estimated Parameters Values	Model Interpretation
$R = \beta_0 + \beta_1 \times CTRP + \beta_2 \times P + \varepsilon$	$R = 41008.84 + 5931.85 CTRP + 3.136 \times P$	Variation in $R$ explained by $CTRP$ and $P$
$R = \alpha_0 + \alpha_1 \times CTRP + \varepsilon_1$	$R = 625763.106 + 4569.214 \times CTRP$	Variation in $R$ explained by $CTRP$
$P = \delta_0 + \delta_1 \times CTRP + \varepsilon_2$	$P = 186456.659 - 434.495 \times CTRP$	Variation in $P$ explained by $CTRP$
$\varepsilon_1 = \eta_0 + \eta_1 \times \varepsilon_2$	$\varepsilon_1 = 3.136 \times \varepsilon_2$ (The value of $\eta_1$ is same as that of $\beta_2$ )	$\varepsilon_1$ is variation in $R$ not explained by $CTRP$ $\varepsilon_2$ is variation in $P$ not explained by $CTRP$

That is, every new variable added to the model is partialled out from other independent variables and regressed with the partialled out dependent variable. Regression model [Eq. (10.25)] is between partialled out dependent variable (Revenue) and partialled out promotion expenditure.

The partial regression coefficient provides the change in the response variable for a unit change in the explanatory variable, when all other explanatory variables are kept constant or controlled. For example, in regression model [Eq. (10.22)], for every one unit increase in  $CTRP$ , the revenue increases by 5931.84 provided the promotion expenses are kept constant. Similarly when the promotion is increased by one unit, the revenue increases by 3.136 provided  $CTRP$  is kept constant. However, in practice, it may not be possible to control a variable in many situations.

## 10.6 | STANDARDIZED REGRESSION CO-EFFICIENT

One frequently asked question in multiple regression models is the impact of different explanatory variables on the response variable. For example, for the model defined in Eq. (10.21), the coefficient value for  $CTRP$  is 5931.85 and the coefficient for promotion spend is 3.136. However, this does not mean that  $CTRP$  has more influence on the revenue compared to promotion expenses,  $P$ . The reason is that the unit of measurement for  $CTRP$  is different from the unit of measurement of  $P$ . We have to derive standardized regression coefficients to compare the impact of different explanatory variables that have different units of measurement.

Since the regression coefficients cannot be compared directly due to difference in scale and units of measurements of variables, one has to normalize the data to compare the regression coefficients and their impact on the response variable. A regression model can be built on standardized dependent variable and standardized independent variables, the resulting regression coefficients are then known as **standardized regression coefficients**.

The standardized regression coefficient (standardized beta) can also be calculated using the following formula:

$$\text{Standardized Beta} = \hat{\beta} \times \left( \frac{S_{X_i}}{S_Y} \right) \quad (10.26)$$

where  $S_{X_i}$  is the standard deviation of the explanatory variable  $X_i$  and  $S_Y$  is the standard deviation of the response variable  $Y$ . For the data in Table 10.1, the standard deviations are given by

$$S_Y = 136527.88, \quad S_{CTRP} = 16.85, \quad S_P = 32052.62$$

Standardized regression coefficient for  $CTRP = 5931.85 \times (16.85/136527.88) = 0.732$

Standardized regression coefficient for  $P = 3.136 \times (32052.62/136527.88) = 0.736$

The standardized regression coefficients can be interpreted as follows: For one standard deviation change in the explanatory variable, the standard regression coefficient captures the number of standard deviations by which the response variable will change. For example, when  $CTRP$  is changed by one standard deviation,  $Y$  will change by 0.732 standard deviations. Similarly, when  $P$  changes by one standard deviation,  $Y$  will change by 0.736 standard deviations. That is, the variable  $P$  has slightly higher impact on the revenue compared to  $CTRP$ . The SPSS output for Example 10.1 with standardized regression coefficient is shown in Table 10.11.

**TABLE 10.11** Standardized regression coefficients

Model	Unstandardized Coefficients		Standardized Coefficients	<i>t</i>	Sig.
	<i>B</i>	Std. Error	Beta		
1	(Constant)	41008.840	90958.920	0.451	0.655
	<i>CTRP</i>	5931.850	576.622	10.287	0.000
	<i>P</i>	3.136	0.303	10.344	0.000

## 10.7 | REGRESSION MODELS WITH QUALITATIVE VARIABLES

In MLR, many predictor variables are likely to be qualitative or categorical variables. Since the scale is not a ratio or interval for categorical variables, we cannot include them directly in the model, since its inclusion directly will result in model misspecification. We have to pre-process the categorical variables using dummy variables for building a regression model. For example, consider a regression model between dependent variable  $Y$  and an independent marital status ( $MS$ ). It will be incorrect to use the functional form in Eq. (10.26) to find the relationship  $Y$  and marital status.

$$Y = \beta_0 + \beta_1 MS + \varepsilon \quad (10.27)$$

In the data, various marital status are probably captured using a code (such as 1, 2, etc.). Assume that the following codes are used in the data to describe various categories of marital status:

- $MS = 1$  implies marital status “single”
- $MS = 2$  implies marital status “married”
- $MS = 3$  implies marital status “divorced”

Equation (10.26) will then become

$$\begin{aligned} Y &= \beta_0 + \beta_1 \text{ for } MS = 1 \\ Y &= \beta_0 + 2\beta_1 \text{ for } MS = 2 \\ Y &= \beta_0 + 3\beta_1 \text{ for } MS = 3 \end{aligned}$$

Assuming  $\beta_0 = 0$ , the value of  $Y$  for divorced is 3 times the value of  $Y$  for single! In other words, when  $\beta_0 = 0$

$$\frac{E(Y | X = 3)}{E(Y | X = 1)} = \frac{3\beta_1}{1\beta_1} = 3$$

That is, the value of dependent variable  $Y$  for marital status divorced is 3 times better (or worse when  $\beta_1$  is negative) than the value of dependent variable for marital status single. We will reach such incorrect conclusion which is due to that fact that the model in Eq. (10.27) is incorrect (known as model misspecification). This comes from the fact that the values associated with different categories of marital status are just labels and it is not a ratio or interval scale. Whenever we have a categorical variable in the data, we have to convert them into dummy variables before including them in a regression model. The correct model specification is

$$Y = \beta_0 + \beta_1 S + \beta_2 M + \varepsilon \quad (10.28)$$

where  $S$  and  $M$  are dummy variables created for marital status category single and married, respectively. Depending on the marital status of a specific observation, Eq. (10.28) will have the following three cases:

$$\begin{aligned} Y &= \beta_0 \text{ (when the marital status is divorced)} \\ Y &= \beta_0 + \beta_1 \text{ (when the marital status is single)} \\ Y &= \beta_0 + \beta_2 \text{ (when the marital status is married)} \end{aligned}$$

Whenever, we have  $n$  levels (or categories) for a qualitative variable (categorical variable), we will use  $(n - 1)$  dummy variables, where each dummy variable is a binary variable used for representing whether an observation belongs to a category or not. That is, we have to pre-process the data to represent the categorical variables using dummy variables. The reason why we create only  $(n - 1)$  dummy variables is that inclusion of dummy variables for all categories and the constant in the regression equation will create perfect multi-collinearity (will be discussed later) and the matrix  $X$  in Eq. (10.2) will become singular.

**IMPORTANT**

If we create dummy variables for all categories and also include the constant in the model, it will create perfect multi-collinearity and the matrix  $X$  in Eq. (10.2) will become a singular matrix [since the summation all the dummy variable will add to 1 and the coefficient corresponding to constant  $\beta_0$  will be 1 in matrix  $X$  of Eq. (10.2)].

**EXAMPLE 10.2**

The data in Table 10.12 provides salary and educational qualifications of 30 randomly chosen people in Bangalore. Build a regression model to establish the relationship between salary earned and their educational qualifications.

Note that, if we build a model  $Y = \beta_0 + \beta_1 \times \text{Education}$ , it will be incorrect. We have to use 3 dummy variables since there are 4 categories for educational qualification. Data in Table 10.12 has to be pre-processed using 3 dummy variables (*HS*, *UG*, and *PG*) as shown in Table 10.13.

**TABLE 10.12** Education versus salary

S. No.	Education <sup>a</sup>	Salary	S. No.	Education	Salary	S. No.	Education	Salary
1	1	9800	11	2	17200	21	3	21000
2	1	10200	12	2	17600	22	3	19400
3	1	14200	13	2	17650	23	3	18800
4	1	21000	14	2	19600	24	3	21000
5	1	16500	15	2	16700	25	4	6500
6	1	19210	16	2	16700	26	4	7200
7	1	9700	17	2	17500	27	4	7700
8	1	11000	18	2	15000	28	4	5600
9	1	7800	19	3	18500	29	4	8000
10	1	8800	20	3	19700	30	4	9300

<sup>a</sup> 1 – High school, 2 – Under-graduate, 3 – Post-graduate and 4 – None.

**TABLE 10.13** Pre-processed data (sample)

Observation	Education	Pre-Processed data			Salary
		High School (HS)	Under-Grauate (UG)	Post-Graduate (PG)	
1	1	1	0	0	9800
11	2	0	1	0	17200
19	3	0	0	1	18500
27	4	0	0	0	7700

The corresponding regression model is as follows:

$$Y = \beta_0 + \beta_1 \times HS + \beta_2 \times UG + \beta_3 \times PG \quad (10.29)$$

where HS, UG, and PG are the dummy variables corresponding to the categories high school, under-graduate, and post-graduate, respectively. The fourth category (none) for which we did not create an explicit dummy variable is called the base category. In Eq. (10.29), when  $HS = UG = PG = 0$ , the value of  $Y$  is  $\beta_0$ , which corresponds to the education category, ‘none’. The SPSS output for the regression model in Eq. (10.29) using the data in Table 10.12 is shown in Table 10.14.

**TABLE 10.14** Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients		<i>t</i> -value	<i>p</i> -value
	<i>B</i>	Std. Error	Beta			
1	(Constant)	7383.333	1184.793		6.232	0.000
	High-School (HS)	5437.667	1498.658	0.505	3.628	0.001
	Under-Graduate (UG)	9860.417	1567.334	0.858	6.291	0.000
	Post-Graduate (PG)	12350.000	1675.550	0.972	7.371	0.000

The corresponding regression equation is given by

$$Y = 7383.33 + 5437.667 \times HS + 9860.417 \times UG + 12350.00 \times PG$$

Note that in Table 10.4, all the dummy variables are statistically significant at  $\alpha = 0.01$ , since *p*-values are less than 0.01.

### 10.7.1 | Interpretation of Regression Coefficients of Categorical Variables

In regression model with categorical variables, the regression coefficient corresponding to a specific category represents the change in the value of  $Y$  from the base category value ( $\beta_0$ ). For example, in Eq. (10.28), when  $HS = UG = PG = 0$ , the value of  $Y = 7383.33$ . In this case, the base category is the education ‘none’. That is, when education category is none, the average salary is 7383.33. When education category is ‘HS’, we get

$$Y = 7383.333 + 5437.667 = 12821.00$$

That is, 5437.667 is the shift or deviation from the base category for category ‘HS’ (education category high school). Note that this interpretation is possible only when the dummy variable corresponding to a category is **statistically significant** (that is, the corresponding *p*-value is less than the significance value  $\alpha$ ). It is possible that few dummy variables may not be statistically significant in a regression model with categorical variable. If a dummy variable is not significant then it implies that it is not statistically different from the base category since we retain the null hypothesis (there is no relationship) in *t*-test corresponding to that dummy variable. In such a case, we combine the category with the base category and develop the model again using new set of dummy variables.

**IMPORTANT**

If a dummy variable is not statistically significant (that is the  $p$ -value is greater than  $\alpha$ ) then it implies that the category corresponding to that dummy variable is not statistically different from the base category. We have to then combine the category with the base category in order to create a new set of dummy variables and develop the model again.

### 10.7.2 | Interaction Variables in Regression Models

Interaction variables are basically inclusion of variables in the regression model that are a product of two independent variables (such as  $X_1 X_2$ ). Usually the interaction variables are product of a continuous and a categorical variable. The inclusion of interaction variables enables the data scientists to check the existence of conditional relationship between the dependent variable and two independent variables. For example, consider the relation between salary ( $Y$ ), gender ( $X_1$ ), work experience ( $X_2$ ), and interaction between gender and work experience ( $X_1 X_2$ ) as shown below:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \quad (10.30)$$

In the regression equation (10.30), the term  $X_1 X_2$  is an interaction variable. Assume that female is coded as 1 and male as 0 for variable gender ( $X_1$ ) in the data. For female and male, Eq. (10.30) can be written as

For  $X_1 = 1$  (Female)

$$Y = \beta_0 + \beta_1 + (\beta_2 + \beta_3) X_2 \quad (10.31)$$

For  $X_1 = 0$  (Male)

$$Y = \beta_0 + \beta_2 X_2 \quad (10.32)$$

If the regression coefficients  $\beta_2$  and  $\beta_3$  are statistically significant, then the rate at which the value of  $Y$  changes for change in  $X_2$  for male and female will be different. That is, the rate at which  $Y$  changes for female as  $X_2$  changes is  $(\beta_2 + \beta_3)$  whereas as the rate of change of  $Y$  for male when  $X_2$  changes is  $\beta_2$ . In other words, the relation between salary  $Y$  and work experience  $X_2$  is conditioned on the variable  $X_1$ . That is, the regression model will result in two different equations as shown in Eq. (10.31) and Eq. (10.32).

#### EXAMPLE 10.3

The data in Table 10.15 provides salary, gender, and work experience (WE) of 30 workers in a firm. In Table 10.15, gender = 1 denotes female and 0 denotes male and WE is the work experience in number of years. Build a regression model by including an interaction variable between gender and work experience. Discuss the insights based on the regression output.

**TABLE 10.15** Data on salary, gender, and work experience (WE)

S. No.	Gender	WE	Salary	S. No.	Gender	WE	Salary
1	1	2	6800	16	0	2	22100
2	1	3	8700	17	0	1	20200
3	1	1	9700	18	0	1	17700
4	1	3	9500	19	0	6	34700
5	1	4	10100	20	0	7	38600
6	1	6	9800	21	0	7	39900
7	0	2	14500	22	0	7	38300
8	0	3	19100	23	0	3	26900
9	0	4	18600	24	0	4	31800
10	0	2	14200	25	1	5	8000
11	0	4	28000	26	1	5	8700
12	0	3	25700	27	1	3	6200
13	0	1	20350	28	1	3	4100
14	0	4	30400	29	1	2	5000
15	0	1	19400	30	1	1	4800

**Solution:**

Let the regression model be

$$Y = \beta_0 + \beta_1 \times \text{Gender} + \beta_2 \times \text{WE} + \beta_3 \times \text{Gender} \times \text{WE}$$

The SPSS output for the regression model including interaction variable is given in Table 10.16.

**TABLE 10.16** Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1	(Constant)	13443.895	1539.893	8.730	0.000
	Gender	-7757.751	2717.884	-0.348	-2.854 0.008
	WE	3523.547	383.643	0.603	9.184 0.000
	Gender*WE	-2913.908	744.214	-0.487	-3.915 0.001

The regression equation is given by

$$Y = 13443.895 - 7757.751 \text{ Gender} + 3523.547 \text{ WE} - 2913.908 \text{ Gender} \times \text{WE} \quad (10.33)$$

Equation (10.33) can be written as

For Female (Gender = 1)

$$Y = 13443.895 - 7757.751 + (3523.547 - 2913.908) \text{ WE} \quad (10.34)$$

For Male (Gender = 1)

$$Y = 13443.895 + 3523.547 \text{ WE} \quad (10.35)$$

That is, the change in salary for female when WE increases by one year is 609.639 and for male the increase is 3523.547. That is, the salary for male workers is increasing at a higher rate compared female workers. Interaction variables are an important class of derived variables in regression model building.

## 10.8 | VALIDATION OF MULTIPLE REGRESSION MODEL

The following measures and tests are carried out to validate a multiple linear regression model:

1. Coefficient of multiple determination ( $R$ -Square) and Adjusted  $R$ -Square, which can be used to judge the overall fitness of the model.
2.  $t$ -test to check the existence of statistically significant relationship between the response variable and individual explanatory variable at a given significance level ( $\alpha$ ) or at  $(1 - \alpha)100\%$  confidence level.
3.  $F$ -test to check the statistical significance of the overall model at a given significance level ( $\alpha$ ) or at  $(1 - \alpha)100\%$  confidence level.
4. Conduct a residual analysis to check whether the normality, homoscedasticity assumptions have been satisfied. Also, check for any pattern in the residual plots to check for correct model specification.
5. Check for presence of multi-collinearity (strong correlation between independent variables) that can destabilize the regression model.
6. Check for auto-correlation in case of time-series data.

It is important that the aforementioned tests are conducted before the model is selected for deployment.

## 10.9 | CO-EFFICIENT OF MULTIPLE DETERMINATION ( $R$ -SQUARE) AND ADJUSTED $R$ -SQUARE

As in the case of simple linear regression,  $R$ -square measures the proportion of variation in the dependent variable explained by the model. The co-efficient of multiple determination ( $R$ -Square or  $R^2$ ) is given by

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^N (\hat{Y}_i - Y_i)^2}{\sum_{i=1}^N (Y_i - \bar{Y})^2} \quad (10.36)$$

In Eq. (10.36), SSE is the sum of squares of errors and SST is the sum of squares of total deviation and  $N$  is the sample size. In case of MLR, SSE will decrease as the number of explanatory variables increases, and SST remains constant. So, it is possible, that  $R$ -square will increase even when there is no statistically significant relationship between the explanatory variable and the response variable. To counter this,

$R^2$  value is adjusted by normalizing both  $SSE$  and  $SST$  with the corresponding degrees of freedom. The adjusted  $R$ -square with  $k$  predictors is given by

$$\text{Adjusted } R\text{-Square} = 1 - \frac{SSE / (N - k - 1)}{SST / (N - 1)} \quad (10.37)$$

While  $R$ -square is a non-decreasing function, adjusted  $R$ -square is not. So, when a new variable is added, it is worth checking whether there is any increase in adjusted  $R^2$  than  $R^2$ .  $R$ -Square and adjusted  $R$ -square values for Example 10.1 are 0.832 and 0.822, respectively (Table 10.2). The adjusted  $R$ -square value is always less than or equal to the  $R$ -square value. No increase in adjusted  $R$ -square after adding a new predictor variable to the model may indicate that the newly added variable may not be statistically significant or it is not explaining the variation in the response variables that is not explained by the variables that are already present in the model.

## 10.10 | STATISTICAL SIGNIFICANCE OF INDIVIDUAL VARIABLES IN MLR – $t$ -TEST

Checking the statistical significance of individual variables is achieved through  $t$ -test. Note that the estimate of regression coefficient is given by Eq. (10.15):

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

This means the estimated value of regression coefficient is a linear function of the response variable. Since we assume that the residuals follow normal distribution,  $Y$  follows a normal distribution and the estimate of regression coefficient also follows a normal distribution. Since the standard deviation of the regression coefficient is estimated from the sample, we use a  $t$ -test. The null and alternative hypotheses in the case of individual independent variable and the dependent variable  $Y$  is given, respectively, by

$H_0$ : There is no relationship between independent variable  $X_i$  and dependent variable  $Y$

$H_A$ : There is a relationship between independent variable  $X_i$  and dependent variable  $Y$

Alternatively,

$H_0$ :  $\beta_i = 0$

$H_A$ :  $\beta_i \neq 0$

The corresponding test statistic is given by

$$t = \frac{\hat{\beta}_i - 0}{S_e(\hat{\beta}_i)} = \frac{\hat{\beta}_i}{S_e(\hat{\beta}_i)} \quad (10.38)$$

For example (10.1), the values of  $t$ -statistic for the variables  $CTRP$  and  $Promotion$ ,  $P$ , are 10.287 and 10.344, respectively (from Table 10.3). The  $t$ -critical value for  $\alpha = 0.05$  and  $df = 35$  is 2.0301. Since the  $t$ -statistic value is much higher than the critical value, we reject the null hypothesis. The corresponding  $p$ -values for  $CTRP$  and  $P$  are close to 0 (Table 10.3).

## 10.11 | VALIDATION OF OVERALL REGRESSION MODEL: $F$ -TEST

Analysis of Variance (ANOVA) is used to validate the overall regression model. If there are  $k$  independent variables in the model, then the null and the alternative hypotheses are, respectively, given by

$$H_0: \beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0$$

$H_1$ : Not all  $\beta$ s are zero.

Note that the statement in alternative hypothesis is that “not all  $\beta$ s are zero”, that is, some of those  $\beta$  values may be zero. That is the reason why we have to do the  $t$ -test to check the existence of statistically significant relationship between individual explanatory variables and the response variable.  $F$ -statistic is given by

$$F = MSR/MSE \quad (10.39)$$

For Example 10.1, the ANOVA table is shown in Table 10.17. The  $F$ -statistic value is 86.62 and the corresponding  $p$ -value (significance value) is  $2.793 \times 10^{-14}$ . Since the  $p$ -value is less than 0.05, we reject the null hypothesis. The degrees of freedom for the numerator in Eq. (10.38) is equal to the number of explanatory variables and degrees of freedom for denominator is  $(N - k - 1)$  where  $N$  is the sample size.

**TABLE 10.17** ANOVA table for Example 10.1

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	573761326866.653	2	286880663433.326	86.623	0.000 <sup>b</sup>
	Residual	115913569910.190	35	3311816283.148		
	Total	689674896776.842	37			

## 10.12 | VALIDATION OF PORTIONS OF A MLR MODEL – PARTIAL F-TEST

In many cases, data scientists may like to validate the portions of the model or a subset of explanatory variables. Assume that the data set has  $N$  observations (sample size), and we define two models named ‘full model’ which has  $k$  independent variables and ‘reduced model’ which has  $r$  independent variables ( $r < k$ ) defined as follows:

Full Model (with all  $k$  explanatory variables):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k \quad (10.40)$$

Reduced Model (with  $r$  explanatory variables, where  $r < k$ ):

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \dots + \alpha_r X_r \quad (10.41)$$

The objective of the partial  $F$ -test is to check where the additional variables ( $X_{r+1}, X_{r+2}, \dots, X_k$ ) in the full model are statistically significant.

The corresponding partial  $F$ -test has the following null and alternative hypotheses:

$$H_0: \beta_{r+1} = \beta_{r+2} = \dots = \beta_k = 0$$

$H_1$ : Not all  $\beta_{r+1}, \beta_{r+2}, \dots, \beta_k$  are zero

The partial  $F$ -test statistic is given by

$$\text{Partial } F = \left( \frac{(SSE_R - SSE_F) / (k - r)}{MSE_F} \right) \quad (10.42)$$

$SSE_R$  is the sum of squared errors of reduced model [Eq. (10.41)],  $SSE_F$  is the sum of squared error of full model [Eq. (10.40)],  $MSE_F$  is the mean squared error of the full model [Eq. (10.40)], and  $(k - r)$  is the difference between the number of variables between full model and the reduced model. Equation (10.42) is equivalent to the following equation:

$$\text{Partial } F = \frac{(R_{\text{full}}^2 - R_{\text{reduced}}^2) / (k - r)}{(1 - R_{\text{full}}^2) / (N - k - 1)} \quad (10.43)$$

where  $R_{\text{full}}^2$  is the coefficient of determination for the full model and  $R_{\text{reduced}}^2$  is the coefficient of determination for the reduced model.

Partial  $F$ -test is used for variable selection in stepwise regression models. That is, if there are several variables and we would like to select one variable at a time to build the model, then partial  $F$ -test can be used for the selection of the variable. We will be discussing variable selection and the stepwise regression model in Section 10.17.

### 10.13 | RESIDUAL ANALYSIS IN MULTIPLE LINEAR REGRESSION

Residual analysis is important for checking assumptions about normal distribution of residuals, homoscedasticity, and the functional form of a regression model. The normal probability plot and residual plot for Example 10.1 are shown in Figures 10.3 and 10.4. If the residuals do not follow normal distribution, then we cannot trust the  $p$ -values of  $t$ -test and  $F$ -test since for the statistic to follow  $t$ -distribution and  $F$ -distribution, the residuals should follow normal or approximate normal distribution. There are many reasons why residuals may not be normal; one such case is misspecification of functional form of regression, that is, the data scientist may have used linear model instead of log-linear or log-log model.

Based on Figures 10.3 and 10.4, one can infer that the residuals in Table 10.1 follow approximate normal distribution and the residual plot reveals that there is no evidence for heteroscedasticity.

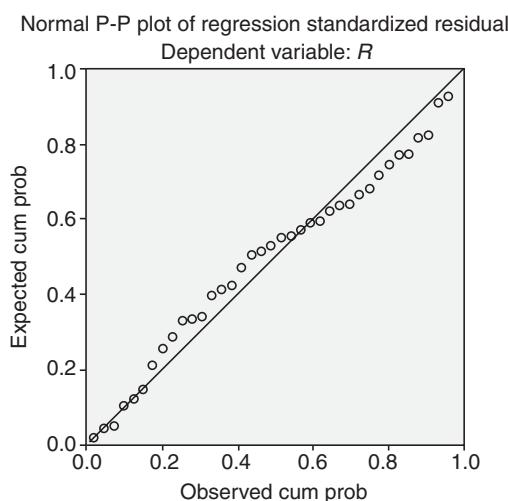
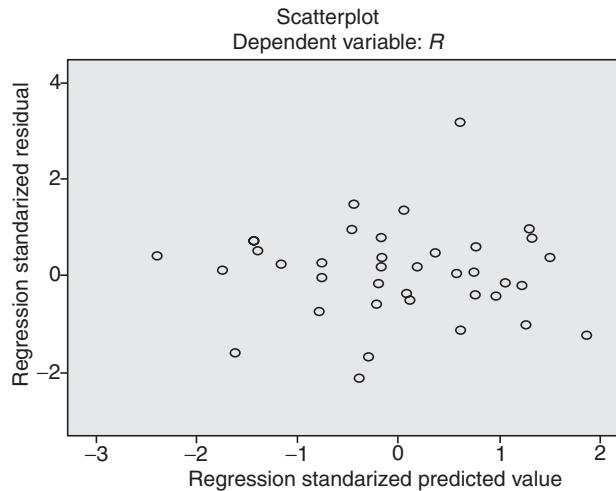


FIGURE 10.3 Normal probability plot (P-P Plot) for Example 10.1.



**FIGURE 10.4** Residual plot for Example 10.1.

## 10.14 | MULTI-COLLINEARITY AND VARIANCE INFLATION FACTOR

When the data set has a large number of independent variables, it is possible that few of these independent variables may be highly correlated. Existence of high correlation between independent variables is called multi-collinearity. Presence of multi-collinearity can destabilize the multiple regression model. Thus, it is necessary to identify the presence of multi-collinearity and take corrective actions. Multi-collinearity can have the following impact on the model:

1. The standard error of estimate of a regression coefficient may be inflated, and may result in retaining of null hypothesis in  $t$ -test, resulting in rejection of a statistically significant explanatory variable. The  $t$ -statistic value is  $(\hat{\beta}/S_e(\hat{\beta}))$ . If  $S_e(\hat{\beta})$  is inflated, then the  $t$ -value will be underestimated resulting in high  $p$ -value that may result in failing to reject the null hypothesis. Thus, it is possible that a statistically significant explanatory variable may be labelled as statistically insignificant due to the presence of multi-collinearity.
2. The sign of the regression coefficient may be different, that is, instead of negative value for regression coefficient, we may have a positive regression coefficient and vice versa.
3. Adding/removing a variable or even an observation may result in large variation in regression coefficient estimates.

### 10.14.1 | Variance Inflation Factor (VIF)

Variance inflation factor (VIF) measures the magnitude of multi-collinearity. Let us consider a regression model with two explanatory variables defined as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \quad (10.44)$$

To find whether there is multi-collinearity, we develop a regression model between the two explanatory variables as follows:

$$X_1 = \alpha_0 + \alpha_1 X_2 \quad (10.45)$$

Let  $R_{12}^2$  be the  $R$ -square value for the regression model in Eq. (10.45). Note that, when there are only two independent variables, we can use correlation coefficient to calculate  $R^2$  between the two variables. Variance inflation factor (*VIF*) is then given by

$$VIF = \frac{1}{1 - R_{12}^2} \quad (10.46)$$

The value  $1 - R_{12}^2$  is called the **tolerance**.  $\sqrt{VIF}$  is the value by which the standard error of estimate is inflated in the presence of multi-collinearity, or  $\sqrt{VIF}$  is the value by which the  $t$ -statistic is deflated. So, the actual  $t$ -value is given by

$$t_{\text{actual}} = \left( \frac{\hat{\beta}_1}{S_e(\hat{\beta}_1)} \right) \times \sqrt{VIF} \quad (10.47)$$

There will be some correlation between explanatory variables in almost all cases, thus the value of *VIF* is likely to be more than one. The data scientists have to decide when to intervene based on the *VIF* value. The **threshold value** for *VIF* is 4 (a few authors suggest 10). *VIF* value of greater than 4 requires further investigation to assess the impact of multi-collinearity. Before building the multiple regression models, it is advised to check the correlation between different explanatory variables for potential multi-collinearity. *VIF* value equal to 4 implies that the  $t$ -statistic value is deflated by a factor 2 and thus there will be a significant increase in the corresponding  $p$ -value. The serious impact of multi-collinearity is that it can change the sign of the regression coefficient (for example, instead of positive, the model may have negative regression coefficient for a predictor and vice versa).

### 10.14.2 | Remedies for Handling Multi-Collinearity

There are many approaches that can be used to handle the impact of multi-collinearity. One easier approach is to remove one of the variables from the model building. For example, the data scientist may remove a variable that is either difficult or expensive to collect. Another approach suggested by researchers is to use centered variables, that is, use  $(X_i - \bar{X}_i)$  instead of  $X_i$ . When there are many variables in the data, the data scientists can use **Principle Component Analysis** (PCA) to avoid multi-collinearity. PCA will create orthogonal components and thus remove potential multi-collinearity. In the recent years, authors use advanced regression models such as **Ridge regression** and **LASSO regression** to handle multi-collinearity.

## 10.15 | AUTO-CORRELATION

Auto-correlation is the correlation between successive error terms in a time-series data. Consider a time-series model as defined below:

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t \quad (10.48)$$

In the regression model [Eq. (10.48)], the values of the response variable  $Y$  are measured at different time points  $t$  and  $X_t$  is the value of the independent variable at time  $t$ . One of the assumptions of regression model is that, there should be no correlation between error terms,  $\varepsilon_t$  and  $\varepsilon_{t-1}$  (known as **auto-correlation of errors of lag 1**). In general, errors  $\varepsilon_t$  and  $\varepsilon_{t-k}$  may be correlated (known as **auto-correlation of lag k**). If there is an auto-correlation, the standard error estimate of the beta coefficient may be underestimated and that will result in overestimation of the  $t$ -statistic value, which, in turn, will result in a low  $p$ -value. Thus, a variable which has no statistically significant relationship with the response variable may be accepted in the model due to the presence of auto-correlation. The presence of auto-correlation can be established using **Durbin–Watson test**.

### 10.15.1 | Durbin–Watson Test for Auto-Correlation

Durbin–Watson is a hypothesis test to check the existence of auto-correlation (Durbin and Watson, 1950). Let  $\rho$  be the correlation between error terms ( $\varepsilon_t, \varepsilon_{t-1}$ ). The null and alternative hypotheses are stated below:

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

The Durbin–Watson statistic,  $D$ , for correlation between errors of one lag is given by

$$D = \frac{\sum_{i=2}^N (e_i - e_{i-1})^2}{\sum_{i=1}^N e_i^2} \cong 2 \left( 1 - \frac{\sum_{i=2}^N e_i e_{i-1}}{\sum_{i=1}^N e_i^2} \right) \quad (10.49)$$

The value of  $D$  statistic in Eq. (10.49) will lie between 0 and 4. The Durbin–Watson test has two critical values,  $D_L$  and  $D_U$ . The inference of the test can be made based on the following conditions (Gujarati and Sangeetha, 2010):

1. If  $D < D_L$ , then the errors are positively correlated.
2. If  $D > D_U$ , then there is no evidence for positive auto-correlation.
3. If  $D_L < D < D_U$ , the Durbin–Watson test is inconclusive.
4. If  $(4 - D) < D_L$ , then errors are negatively correlated.
5. If  $(4 - D) > D_U$ , there is no evidence for negative auto-correlation.
6. If  $D_L < (4 - D) < D_U$ , the test is inconclusive.

As a thumb rule, a Durbin–Watson statistic close to 2 would imply absence of auto-correlation.

### 10.16 | DISTANCE MEASURES AND OUTLIERS DIAGNOSTICS

Outliers can have significant impact on the estimated regression coefficients. That is, the value of regression coefficient may change depending on whether the outliers are present in the data or not. The following distance measures are used for diagnosing the outliers and influential observations in MLR model.

1. Mahalanobis Distance
2. Cook's Distance

3. Leverage Values
4. DFFIT and DFBETA Values

Let us discuss each one of them in detail.

### 10.16.1 | Mahalanobis Distance

As explained in Section 9.7.2, Mahalanobis distance (1936) is a distance between a specific observation and the centroid of all observations of the predictor variables. Here the motivation is that the Euclidian distance will be inappropriate in MLR since variables will have different scales and units of measurements. Thus, Euclidian distance will be meaningless while measuring statistical distances. Mahalanobis distance overcomes the drawbacks of Euclidian distance while measuring distances between multivariate data (Warren *et al.*, 2011). Mahalanobis distance is used for checking outliers among the independent variables. Observations with Mahalanobis distance values of more than chi-square critical value (with degrees of freedom equal to the number of explanatory variables) are classified as outliers.

Let  $X = [X_1, X_2, \dots, X_N]^T$  and  $Y = [Y_1, Y_2, \dots, Y_N]^T$  be the multivariate observations. Mathematically, Mahalanobis distance,  $D_M$ , is given by (Warrant *et al.* 2011)

$$D_M(X_i) = \sqrt{(X_i - \mu_i)S^{-1}(X_i - \mu_i)} \quad (10.50)$$

Here  $D_M(X_i)$  is the Mahalanobis distance of point  $X_i$ ,  $\mu_i$  is the mean, and  $S^{-1}$  is the covariance matrix. The Mahalanobis distance should be less than the chi-square critical value with degrees of freedom equal to the number of independent variables in the model. Alternatively, one can use a thumb rule of 10. Any observation with Mahalanobis distance value of more than 10 is flagged as an **influential observation**.

### 10.16.2 | Cook's Distance

Cook's distance (Cook, 1977) measures the change in the regression parameters and thus how much the predicted value of the dependent variable changes for all the observations in the sample when a particular observation is excluded from sample for the estimation of regression parameters. Cook's distance for multiple linear regression is given by (Bingham, 1977; Chatterjee and Hadi, 1986)

$$D_i = \frac{(\hat{Y}_j - \hat{Y}_{j(i)})^T (\hat{Y}_j - \hat{Y}_{j(i)})}{(k+1) \times MSE} \quad (10.51)$$

where  $D_i$  is the Cook's distance measure for  $i^{\text{th}}$  observation,  $\hat{Y}_j$  is the predicted value of  $j^{\text{th}}$  observation including  $i^{\text{th}}$  observation,  $\hat{Y}_{j(i)}$  is the predicted value of  $j^{\text{th}}$  observation after excluding  $i^{\text{th}}$  observation from the sample,  $k$  is the number of independent variables, and MSE is the mean squared error of the regression model. Cook's distance is a modified Euclidian distance.

A Cook's distance value of more than 1 indicates highly influential observation. Many authors also recommend a value of  $4/(N - k - 1)$  as threshold for Cook's distance (Ryan, 2009). Any value above this will be classified as influential observation and thus requires further investigation.

### 10.16.3 | Leverage Value (or Hat Value)

Leverage value of an observation measures the influence of that observation on the overall fit of the regression function and is related to the Mahalanobis distance. Leverage point  $h_i$  is nothing but the  $i^{\text{th}}$  diagonal element of the hat matrix,  $H = X(X^T X)^{-1} X^T$ . Leverage value for an observation in MLR is given by

$$h_i = [H_{ii}] = X(X^T X)^{-1} X^T \quad (10.52)$$

Using Eqs. (10.50) and (10.52), we can write Mahalanobis distance as (Rousseeuw and Zomeren, 1990)

$$h_i = \frac{\text{Mahalanobis distance}^2}{N-1} + \frac{1}{N} \quad (10.53)$$

For large  $N$ , the leverage value is proportional to Mahalanobis distance [since  $(1/N)$  will be very small]. Leverage value of more than  $2(k+1)/N$  or  $3(k+1)/N$  (Ryan, 2009) is treated as a highly influential observation.

### 10.16.4 | DFFIT AND SDFFIT

DFFIT measures the difference in the fitted value of an observation when that particular observation is removed from the model building. DFFIT is given by

$$\text{DFFIT} = \hat{y}_i - \hat{y}_{i(i)} \quad (10.54)$$

where,  $\hat{Y}_i$  is the predicted value of  $i^{\text{th}}$  observation including  $i^{\text{th}}$  observation,  $\hat{Y}_{i(i)}$  is the predicted value of  $i^{\text{th}}$  observation after excluding  $i^{\text{th}}$  observation from the sample. The standardized DFFIT (SDFFIT) is given by (Belsley *et al.*, 1980; Ryan, 1990)

$$\text{SDFFIT} = \frac{\hat{y}_i - \hat{y}_{i(i)}}{S_e(i)\sqrt{h_i}} \quad (10.55)$$

$S_e(i)$  is the standard error of estimate of the model after removing  $i^{\text{th}}$  observation and  $h_i$  is the  $i^{\text{th}}$  diagonal element in the hat matrix. The threshold for DFFIT is defined using **Standardized DFFIT** (SDFFIT). The absolute value of SDFFIT should be less than  $2\sqrt{(k+1)/N}$ .

### 10.16.5 | DFBETA and SDFBETA

DFBETA measures the change in the regression coefficient when an observation ' $i$ ' is excluded from the model building. DFBETA is given by

$$\text{DFBETA}_i(j) = \hat{\beta}_j - \hat{\beta}_{j(i)} \quad (10.56)$$

where  $\text{DFBETA}_i(j)$  is the change in the regression coefficient for independent variable  $j$  when observation  $i$  is excluded.  $\hat{\beta}_j$  is the estimated value of  $j^{\text{th}}$  regression coefficient including  $i^{\text{th}}$  observation,  $\hat{\beta}_{j(i)}$  is the estimated value of  $j^{\text{th}}$  regression coefficient after excluding  $i^{\text{th}}$  observation from the sample. The standardized DFBETA value (SDFBETA) for observation  $i$  is given by (Belsley *et al.*, 1980; Ryan, 1990)

$$\text{SDFBETA}_i(j) = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{S_e(\hat{\beta}_{j(i)})} \quad (10.57)$$

$\text{SDFBETA}_i(j)$  is the standardized DFBETA value for variable  $j$  after removing observation  $i$  and  $S_e(\hat{\beta}_{j(i)})$  is the standard error of  $\hat{\beta}_j$  after removing observation  $i$ . The threshold for DFBETA is defined using Standardized DFBETA (SDFBETA). The absolute value of SDFBETA should be less than  $2/\sqrt{N}$ . The reason for using SDFFIT and SDFBETA is due to the fact that the change has to be measured in relative terms and not in absolute term and scales of different variables are likely to be different.

Tables 10.18 and 10.19 provide various distance measures obtained from SPSS for Example 10.1.

**TABLE 10.18** Distance measures for data in Example 10.1

S. No.	Mahalanobis Distance	Cook's Distance	Leverage Values	DFFIT	SDFFIT
1	0.48	0.00	0.01	721.10	0.06
2	1.95	0.01	0.05	2299.24	0.14
3	1.17	0.00	0.03	806.00	0.06
4	3.54	0.00	0.10	518.32	0.03
5	0.08	0.02	0.00	2230.54	0.23
6	2.93	0.00	0.08	222.00	0.01
7	2.58	0.00	0.07	2090.11	0.12
8	6.42	0.01	0.17	5024.50	0.19
9	1.40	0.00	0.04	28.16	0.00
10	3.08	0.00	0.08	1191.79	0.06
11	0.87	0.00	0.02	-1323.77	-0.10
12	0.17	0.03	0.00	-3177.81	-0.32
13	2.05	0.05	0.06	-6130.77	-0.38
14	3.47	0.01	0.09	-3316.86	-0.16
15	0.04	0.00	0.00	212.29	0.02
16	1.74	0.03	0.05	-4854.69	-0.31
17	2.60	0.00	0.07	-1262.38	-0.07
18	3.33	0.00	0.09	-595.40	-0.03
19	2.29	0.00	0.06	681.47	0.04
20	3.41	0.02	0.09	4379.00	0.22
21	0.17	0.00	0.00	-424.75	-0.04
22	0.30	0.01	0.01	1826.50	0.17
23	0.28	0.01	0.01	-1361.61	-0.13
24	0.37	0.00	0.01	911.11	0.08

**TABLE 10.18** Distance measures for data in Example 10.1—Continued

S. No.	Mahalanobis Distance	Cook's Distance	Leverage Values	DFFIT	SDFFIT
25	1.18	0.00	0.03	838.11	0.06
26	3.42	0.14	0.09	-12872.78	-0.67
27	5.98	0.16	0.16	-17060.31	-0.69
28	1.63	0.13	0.04	-9532.53	-0.67
29	0.58	0.01	0.02	-1445.23	-0.12
30	0.17	0.01	0.00	1370.87	0.13
31	3.15	0.02	0.09	4718.78	0.24
32	0.98	0.04	0.03	4619.68	0.36
33	1.89	0.02	0.05	3539.23	0.22
34	1.91	0.01	0.05	-2201.26	-0.14
35	0.72	0.01	0.02	-2251.18	-0.18
36	3.83	0.05	0.10	7875.77	0.38
37	2.15	0.00	0.06	-1422.75	-0.08
38	1.69	0.27	0.05	14025.14	1.07

**TABLE 10.19** DFBeta and SDFBeta values for the data in Example 10.1

S. No.	DFB0_1	DFB1_1	DFB2_1	SDFB0_1	SDFB1_1	SDFB2_1
1	531.27	-0.01	8.73	0.01	-0.03	0.01
2	11028.32	-0.03	-52.95	0.12	-0.09	-0.09
3	2233.35	-0.01	-1.39	0.02	-0.04	0.00
4	1800.54	-0.01	-6.46	0.02	-0.02	-0.01
5	-229.00	-0.01	27.31	0.00	-0.03	0.05
6	-528.63	0.00	5.35	-0.01	0.00	0.01
7	-8386.07	0.02	54.81	-0.09	0.05	0.09
8	17048.05	-0.03	-98.81	0.19	-0.10	-0.17
9	-24.45	0.00	-0.15	0.00	0.00	0.00
10	2650.00	-0.02	-1.74	0.03	-0.05	0.00
11	5210.26	-0.01	-40.78	0.06	-0.02	-0.07
12	-11982.58	0.00	69.02	-0.14	0.02	0.12
13	3781.34	-0.08	41.30	0.04	-0.28	0.07
14	-3197.60	-0.03	50.01	-0.03	-0.09	0.09
15	-200.49	0.00	2.58	0.00	0.00	0.00
16	20786.65	-0.04	-133.38	0.23	-0.14	-0.23

(Continued)

**TABLE 10.19** DFBeta and SDFBeta values for the data—Continued

S. No.	DFB0_1	DFB1_1	DFB2_1	SDFB0_1	SDFB1_1	SDFB2_1
17	4067.10	0.00	-34.13	0.04	0.00	-0.06
18	-1603.33	0.00	13.70	-0.02	-0.01	0.02
19	1252.09	0.01	-13.80	0.01	0.02	-0.02
20	-9641.16	-0.01	99.67	-0.11	-0.05	0.17
21	-1342.83	0.00	8.61	-0.01	0.00	0.01
22	6597.90	-0.03	-15.13	0.07	-0.08	-0.03
23	-2120.46	0.02	-8.71	-0.02	0.05	-0.01
24	-848.41	0.01	-1.28	-0.01	0.04	0.00
25	2515.24	-0.01	-2.89	0.03	-0.04	0.00
26	-52562.86	0.07	326.02	-0.60	0.22	0.58
27	31424.69	-0.19	-67.55	0.35	-0.64	-0.12
28	-6900.63	0.12	-102.91	-0.08	0.44	-0.19
29	1603.72	0.01	-31.68	0.02	0.04	-0.05
30	2147.53	-0.01	7.10	0.02	-0.05	0.01
31	18955.65	-0.02	-123.68	0.21	-0.06	-0.21
32	7429.41	-0.07	30.22	0.08	-0.23	0.05
33	-14996.37	0.03	96.65	-0.16	0.10	0.17
34	7723.10	0.00	-63.76	0.08	-0.01	-0.11
35	-10051.24	0.03	33.13	-0.11	0.11	0.06
36	-13029.33	0.10	10.24	-0.14	0.33	0.02
37	3968.00	-0.02	-12.73	0.04	-0.07	-0.02
38	-11709.45	0.20	-76.43	-0.15	0.78	-0.16

The critical values of various distance measures and the corresponding observations from Example 10.1 that require further investigation for the data in Example 10.1 are shown in Table 10.20.

**TABLE 10.20** Influential observations in Example 10.1

S. No.	Distance Measure	Critical Value	Observations that Require Further Investigation
1	Mahalanobis distance	5.99 (chi-square critical value with 2 degrees of freedom)	8
2	Cook's distance	$[4/(N - k - 1)] = 4/35 = 0.1142$	26, 27, 28 and 38
3	Leverage values	$2(k + 1)/N = 6/38 = 0.1578$	8 and 27
4	SDDFIT	$2\sqrt{(k+1)/N} = 2\sqrt{3/38} = 0.5619$	26, 27, 28 and 38
5	SDFBeta	$2/\sqrt{N} = 2/\sqrt{38} = 0.3244$	26, 27, 28 and 38

## 10.17 | VARIABLE SELECTION IN REGRESSION MODEL BUILDING (FORWARD, BACKWARD, AND STEPWISE REGRESSION)

One of the frequently asked questions in multiple regression model development is how to select variables or features to build the model, especially when the data set has a large number of independent variables. In this section we will discuss different approaches that can be used for developing multiple regression models by selecting variables automatically. Such procedures ensure that only statistically significant variables at a significance value of  $\alpha$  are included in the model. There are several criteria used for variable selection to develop MLR models (Beale, 1970; Halinski and Feldt, 1970; Thompson, 1978); in this section we discuss variable selection based on partial  $F$ -test.

### 10.17.1 | Forward Selection

Assume that the data set has  $k$  independent variables. In forward selection, at each step one variable is added to the model. The following steps are used in building regression model using forward selection method.

#### STEP 1

---

Start with no variables in the model. Calculate the correlation between dependent and all independent variables.

---

#### STEP 2

---

Develop simple linear regression model by adding the variable for which the correlation coefficient is highest with the dependent variable (say variable  $X_i$ ). Note that a variable can be added only when the corresponding  $p$ -value is less than the value  $\alpha$ . Let the model be  $Y = \beta_0 + \beta_1 X_i$ . Create a new model  $Y = \alpha_0 + \alpha_1 X_i + \alpha_2 X_j$  ( $j \neq i$ ); there will be  $(k - 1)$  such models. Conduct a partial  $F$ -test to check whether the variable  $X_j$  is statistically significant at  $\alpha$ .

---

#### STEP 3

---

Add the variable  $X_j$  from step 2 with smallest  $p$ -value based on partial  $F$ -test if the  $p$ -value is less than the significance  $\alpha$ .

---

#### STEP 4

---

Repeat step 3 till the smallest  $p$ -value based on partial  $F$ -test is greater than  $\alpha$  or all variables are exhausted.

---

### 10.17.2 | Backward Elimination Procedure

In backward elimination, we enter all independent variables to start with and remove one variable at each step based on partial  $F$ -test.

**STEP 1**

Assume that the data has ' $n$ ' explanatory variables. We start with a multiple regression model with all  $k$  variables. That is,  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$ . We call this full model.

**STEP 2**

Remove one variable at a time repeatedly from the model in step 1 and create a reduced model (say model 2), there will be  $k$  such models. Perform a partial  $F$ -test between the models in step 1 and step 2.

**STEP 3**

Remove the variable with largest  $p$ -value (based on partial  $F$ -test) if the  $p$ -value is greater than the significance  $\alpha$  (or the  $F$ -value is less than the critical  $F$ -value).

**STEP 4**

Repeat the procedure till the  $p$ -value becomes less than  $\alpha$  or there are no variables in the model for which the  $p$ -value is greater than  $\alpha$  based on partial  $F$ -test.

### 10.17.3 | Stepwise Regression

Stepwise regression is a combination of forward selection and backward elimination procedure. In this case, we set the entering criteria ( $\alpha$ ) for a new variable to enter the model based on the smallest  $p$ -value of the partial  $F$ -test and removal criteria ( $\beta$ ) for a variable to be removed from the model if the  $p$ -value exceeds a pre-defined value based on the partial  $F$ -test ( $\alpha < \beta$ ). For example, we may use  $\alpha = 0.05$ . If the smallest  $p$ -value based on the partial  $F$ -test is less than  $\alpha$  then the variable will be entered and if the  $p$ -value is greater than  $\beta = 0.10$ , then we will remove the variable from the equation. At each step a variable is either entered into the model or removed from the model. The first variable to be added to the model is the one that has the highest correlation with the dependent variable provided the  $p$ -value corresponding to that variable is less than the significance value  $\alpha$ .

In Tables 10.21 and 10.22, the stepwise regression model development for the data in Example 10.1 using SPSS is given.

**TABLE 10.21** Variables entered/removed

Model	Variables Entered	Variables Removed	Method
1	$P$	.	Stepwise (Criteria: Probability-of- $F$ -to-enter $\leq 0.050$ , Probability-of- $F$ -to-remove $\geq 0.100$ ).
2	$CTRP$	.	Stepwise (Criteria: Probability-of- $F$ -to-enter $\leq 0.050$ , Probability-of- $F$ -to-remove $\geq 0.100$ ).

From Table 10.21, we can see the first variable that was entered into the model was  $P$  (promotion) followed by  $CTRP$ . The changes in  $R$ -square values are shown in Table 10.22.

**TABLE 10.22** Model summary

Model	$R$	$R$ -Square	Adjusted $R$ -Square	Std. Error of the Estimate
1	0.569	0.324	0.305	113821.793
2	0.912	0.832	0.822	57548.382

From Table 10.22, we know that the  $R$ -square value is 0.324 after entering the 1<sup>st</sup> variable ( $P$ ) and after adding  $CTRP$ , the  $R$ -square increases to 0.832. The difference in  $R^2$  between step 1 and step 2 of the stepwise regression is 0.508, since the increase in  $R^2$  is given by the square of part-correlation (semi-partial correlation). The value of part-correlation between the dependent variable and  $CTRP$  is 0.7127. Table 10.23 gives the regression coefficient values at each step.

**TABLE 10.23** Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients	$t$	Sig.
	$B$	Std. Error	Beta		
1	(Constant)	881383.138	79116.786	11.140	0.000
	$P$	2.424	0.584	4.151	0.000
2	(Constant)	41008.840	90958.920	0.451	0.655
	$P$	3.136	0.303	10.344	0.000
	$CTRP$	5931.850	576.622	10.287	0.000

### 10.18 | AVOIDING OVERFITTING: MALLOWS'S $C_p$

While developing a multiple linear regression model, it is important to identify the ideal number of independent variables in the model. One of the common issues with regression models is over fitting. In the presence of over fitting, the model may perform well in the training data but will perform badly in the test data. This may be caused due to the presence of unnecessary variables in the model. Mallows's  $C_p$  (Mallows, 1973) is used to select the best regression model by incorporating the right number of explanatory variables in the model. Mallow's  $C_p$  is given by

$$C_p = \left( \frac{SSE_p}{MSE_{full}} \right) - (N - 2p) \quad (10.58)$$

where  $SSE_p$  is the sum of squared errors with  $p$  parameters in the model (including constant),  $MSE_{full}$  is the mean squared error with all variables in the model,  $N$  is the number of observations,  $p$  is the number of parameters in the regression model including constant.

The best regression model is the model with number parameters closest to the  $C_p$  value. While developing regression model using variable selection methods such as forward or stepwise, we calculate  $C_p$

after each iteration and the model with number of parameters  $p$  closest to the value  $C_p$  is chosen as the best regression model.

### 10.19 | TRANSFORMATIONS

Transformation is a process of deriving new dependent and/or independent variables to identify the correct functional form of the regression model. For example, the dependent variable  $Y$  may be replaced in the model with  $\ln(Y)$ ,  $\sqrt{Y}$ ,  $1/Y$ , etc. and similarly an independent variable  $X$  may be replaced with  $\ln(X)$ ,  $\sqrt{X}$ ,  $1/X$ , etc. Transformation in MLR is used to address the following issues:

1. Poor fit (low  $R^2$  value).
2. Pattern in residual analysis indicating potential non-linear relationship between the dependent and independent variable. For example,  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$  is used for developing the model instead of  $\ln(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ , resulting in clear pattern in residual plot.
3. Residuals do not follow a normal distribution.
4. Residuals are not homoscedastic.

#### EXAMPLE 10.4

Data on amount of money spent ( $Y$ ) by customers at an e-commerce portal, monthly income ( $X_1$ ), and family size ( $X_2$ ) is collected for 200 customers (File name: Example 10.4.Xlsx). Build a regression models and perform diagnostic tests. Choose the correct regression model.

#### Solution:

We will first develop a model as given below:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

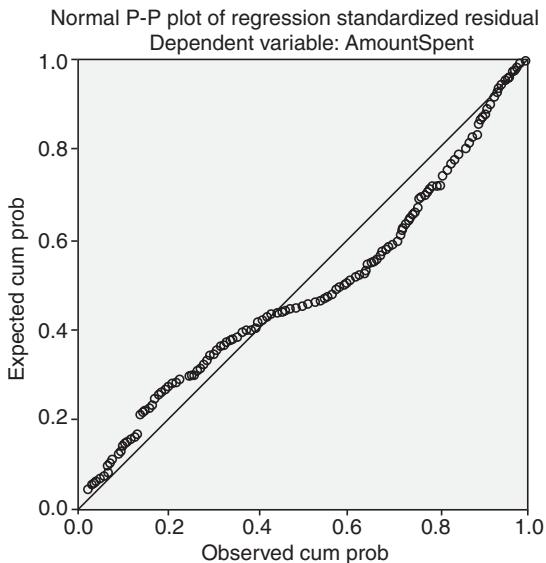
The corresponding SPSS output for model is shown in Tables 10.24 and 10.25. The P-P plot of residual and residual plot between standardized predicted value and standardized residual value are shown in Figures 10.5 and 10.6.

**TABLE 10.24** Model summary

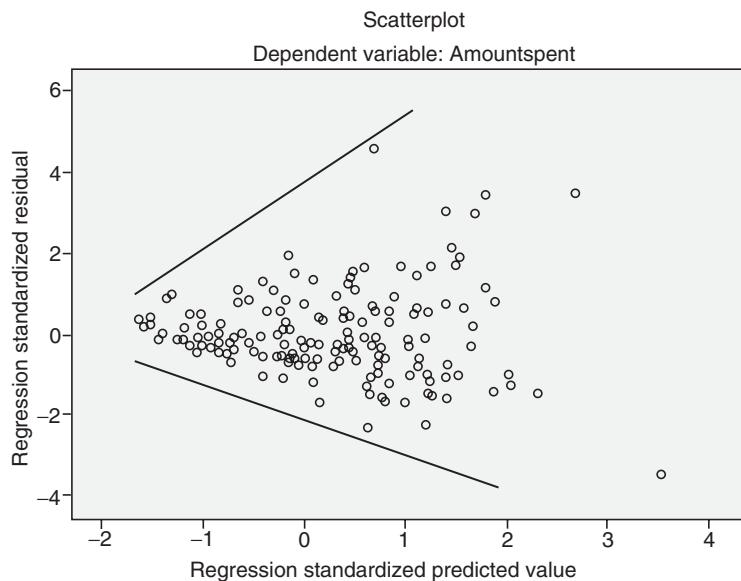
Model	R	R-Square	Adjusted R-Square	Std. Error of the Estimate
1	0.764	0.584	0.580	600.114

**TABLE 10.25** Coefficients

Model	Unstandardized Coefficients			Standardized Coefficients	
	B	Std. Error	Beta	t	Sig.
(Constant)	395.571	137.062		2.886	0.004
1	Income	0.017	0.001	0.752	16.358
	Family Size	-129.415	38.272	-0.155	-3.381
					0.001



**FIGURE 10.5** P-P plot of residuals.



**FIGURE 10.6** Plot between standardized residuals and standardized predicted value.

From Figure 10.5, we know that the residuals do not follow a normal distribution (however, we need only an approximate normal distribution and hence this is not a major concern). Figure 10.6 shows a funnel shape, indicating presence of heteroscedasticity. Since there is heteroscedasticity, the estimate of standard error of

residuals is incorrect and thus  $p$ -values obtained using the statistical tests ( $t$  and  $F$ ) are incorrect.

One reason for heteroscedasticity is the use of an incorrect functional form. One can find the appropriate function form using Box Cox power transformation (1964). Box and Cox (1964) suggested the following transformation to address violation regression assumptions such as normality and homoscedasticity:

$$Y^\lambda = \begin{cases} (Y^\lambda - 1) / \lambda, & \lambda \neq 0 \\ \ln(Y), & \lambda = 0 \end{cases} \quad (10.59)$$

The value of  $\lambda$  can be estimated using maximum likelihood estimation. Alternatively, one can use trial-and-error to find the appropriate functional form. Equation (10.60) is a log-linear functional form for the relationship between the response variable and predictors:

$$\ln(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \quad (10.60)$$

The SPSS outputs for the model in Eq. (10.60) are shown in Tables 10.26 and 10.27. The P-P plot and residual plots are shown in Figures 10.7 and 10.8.

The  $R^2$  for the log-linear model has increased to 0.656 from 0.584 for the linear model. Most importantly, the P-P plot of residuals in Figure 10.7 looks better than the linear model and there is no evidence for heteroscedasticity in Figure 10.8. Thus the log-linear model in Eq. (10.58) is better than linear model in Eq. (10.57).

**TABLE 10.26** Model summary

Model	R	R-Square	Adjusted R-Square	Std. Error of the Estimate
1	0.810	0.656	0.652	0.49932

**TABLE 10.27** Coefficients

Model	Unstandardized Coefficients			Standardized Coefficients		t	Sig.
	B	Std. Error	Beta				
(Constant)	6.109	0.114				53.570	0.000
1	Income	1.665E-005	0.000	0.787		18.812	0.000
	Family Size	-.160	0.032		-0.210	-5.026	0.000

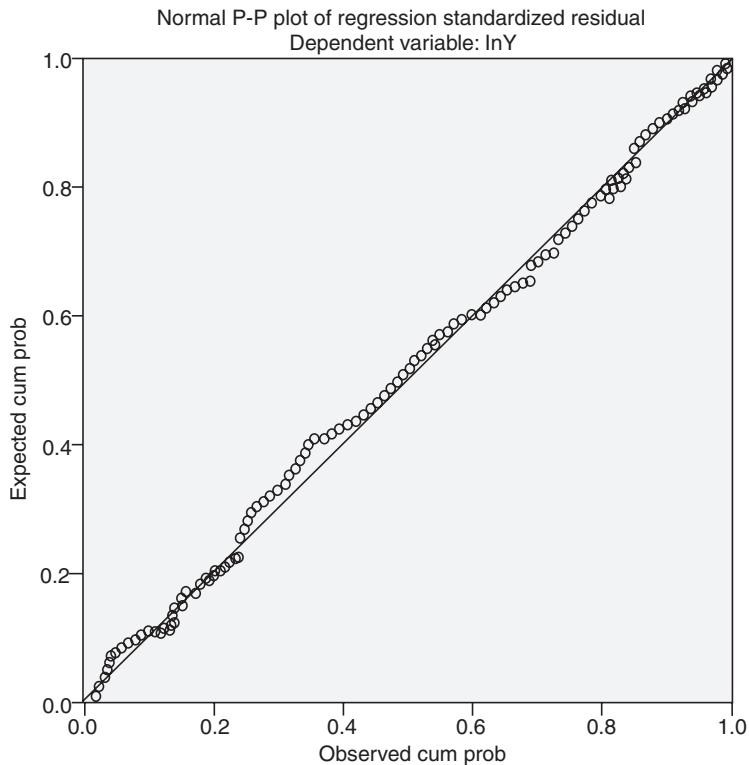


FIGURE 10.7 P-P plot for residuals.

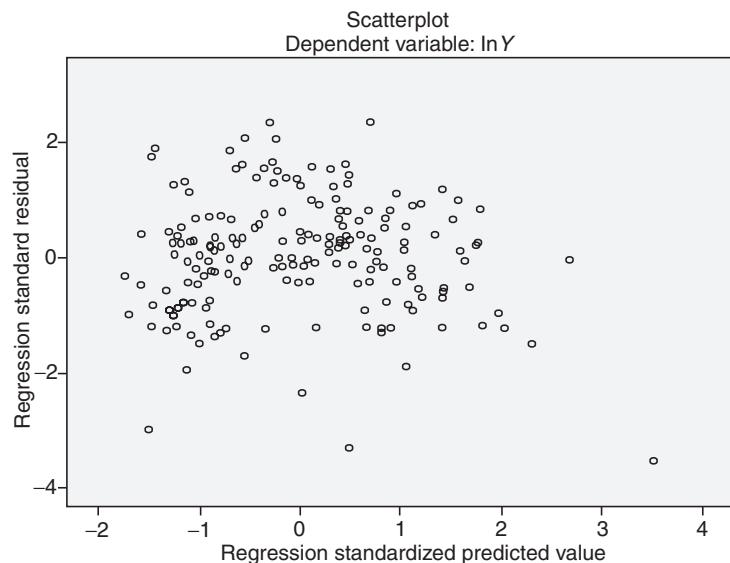


FIGURE 10.8 Plot between standardized residuals and standardized predicted value.

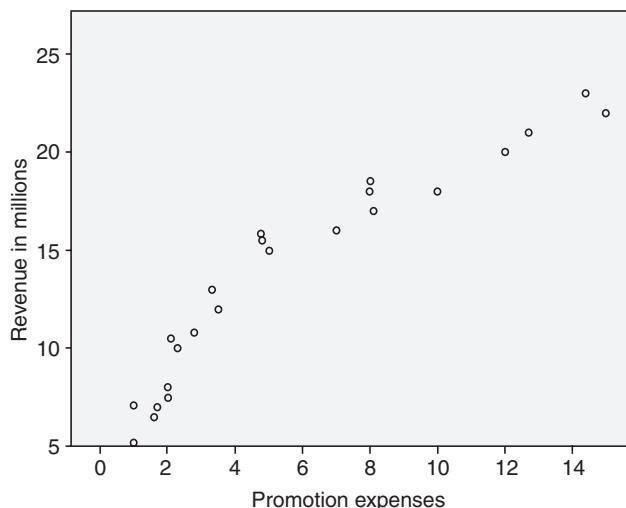
**EXAMPLE 10.5**

Table 10.28 shows the data on revenue generated (in million of rupees) from a product and the promotion expenses (in million of rupees). Develop an appropriate regression model (data file: Example 10.5.Xlsx).

**TABLE 10.28** Data on revenue generated and promotion expenses

S. No.	Revenue in Millions	Promotion Expenses	S. No.	Revenue in Millions	Promotion Expenses
1	5	1	13	16	7
2	6	1.8	14	17	8.1
3	6.5	1.6	15	18	8
4	7	1.7	16	18	10
5	7.5	2	17	18.5	8
6	8	2	18	21	12.7
7	10	2.3	19	20	12
8	10.8	2.8	20	22	15
9	12	3.5	21	23	14.4
10	13	3.3	22	7.1	1
11	15.5	4.8	23	10.5	2.1
12	15	5	24	15.8	4.75

Let  $Y$  = Revenue Generated and  $X$  = Promotion Expenses. The scatter plot between  $Y$  and  $X$  for the data in Table 10.28 is shown in Figure 10.9. It is clear from the scatter plot that the relationship between  $X$  and  $Y$  is not linear; it looks more like a logarithmic function.



**FIGURE 10.9** Scatter plot between promotion expenses and revenue in millions.

Consider the function  $Y = \beta_0 + \beta_1 X$ . The output for this regression is shown in Tables 10.29 and 10.30 and in Figure 10.10. There is a clear increasing and decreasing pattern in Figure 10.10 indicating non-linear relationship between  $X$  and  $Y$ .

**TABLE 10.29** Model summary

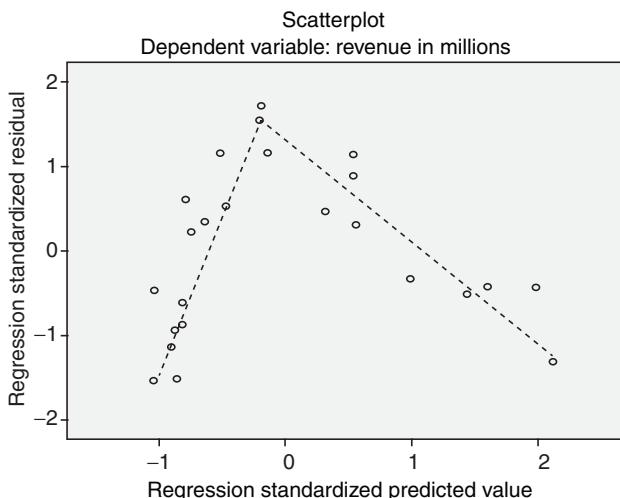
Model	R	R-Square	Adjusted R-Square	Std. Error of the Estimate
1	0.940	0.883	0.878	1.946

**TABLE 10.30** Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients	<i>t</i>	Sig.
	<i>B</i>	Std. Error	Beta		
1	(Constant)	6.831	0.650	10.516	0.000
	Promotion Expenses	1.181	0.091	12.911	0.000

Since there is a pattern in the residual plot, we cannot accept the linear model ( $Y = \beta_0 + \beta_1 X$ ).

Next we try the model  $Y = \beta_0 + \beta_1 \ln(X)$ . The SPSS output for  $Y = \beta_0 + \beta_1 \ln(X)$  is shown in Tables 10.31 and 10.32 and the residual plot is shown in Figure 10.11. Note that for the model  $Y = \beta_0 + \beta_1 \ln(X)$ , the  $R^2$ -value is 0.96 whereas the  $R^2$ -value for the model  $Y = \beta_0 + \beta_1 X$  is 0.883. Most importantly, there is no obvious pattern in the residual plot of the model  $Y = \beta_0 + \beta_1 \ln(X)$ . The model  $Y = \beta_0 + \beta_1 \ln(X)$  is preferred over the model  $Y = \beta_0 + \beta_1 X$ .



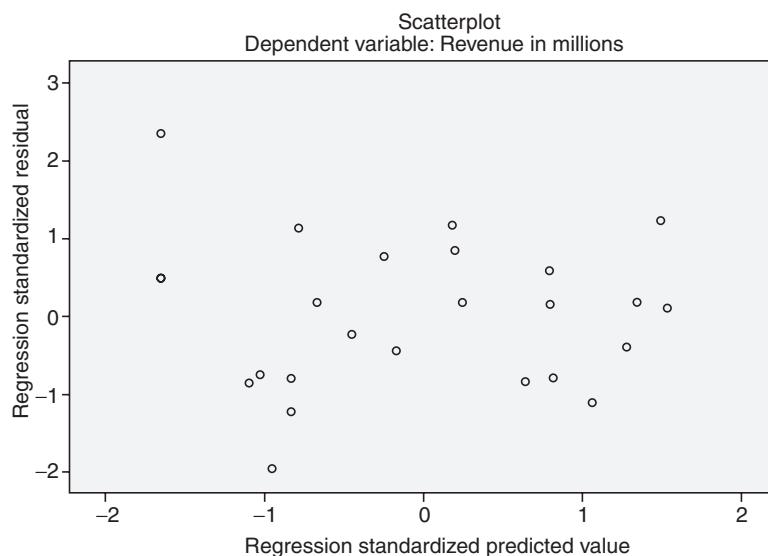
**FIGURE 10.10** Residual plot.

**TABLE 10.31** Model summary

Model	R	R-Square	Adjusted R-Square	Std. Error of the Estimate
1	0.980	0.960	0.959	1.134

**TABLE 10.32** Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients		t	Sig.
	B	Std. Error	Beta			
1	(Constant) 4.439	0.454			9.771	0.000
	In (X) 6.436	0.279	0.980		23.095	0.000

**FIGURE 10.11** Residual plot for the model  $Y = \beta_0 + \beta_1 \ln(X)$ .

### 10.19.1 | Tukey and Mosteller's Bulging Rule for Transformation

Finding the correct functional form of relationship between the dependent variable  $Y$  and independent variables is important in regression model building. In Example 10.4 we transformed  $Y$  to  $\ln(Y)$  to make the residuals homoscedastic and in Example 10.5,  $X$  was transformed to  $\ln(X)$  to remove the pattern in the residual plot. An easier way of identifying an appropriate transformation was provided by Mosteller and Tukey (1977), popularly known as **Tukey's Bulging Rule**. To apply Tukey's Bulging Rule we need to look at the pattern in the scatter plot between the dependent and independent variable. Figure 10.12 shows Tukey's bulging rule which suggests transformations that can be used based on the shape of scatter plot.

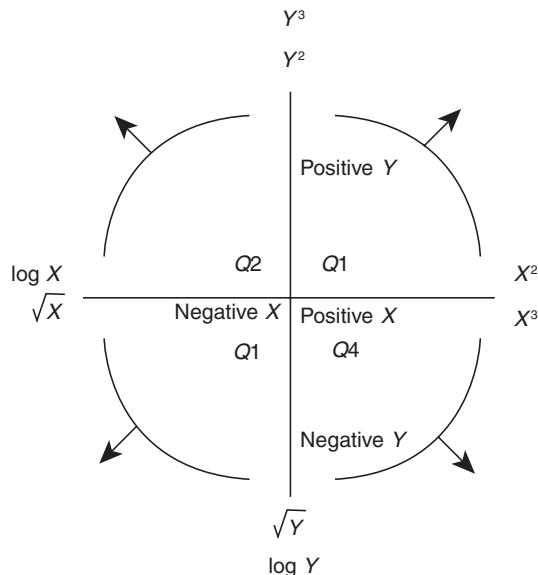


FIGURE 10.12 Tukey's Bulging Rule (adopted from Tukey and Mosteller, 1977).

Table 10.33 provides suggested transformation based on the shape of the scatter plot as identified using the quadrants.

TABLE 10.33 Tukey's rule for transformations

Shape of Scatter Plot	Suggested Transformation for X	Suggested Transformation for Y
Q1 ( $X$ and $Y$ positive)	$X^p$ where $p > 1$ (e.g. $X^2, X^3$ , etc.)	$Y^q$ where $q > 1$ (e.g. $Y^2, Y^3$ , etc.)
Q2 ( $X$ negative and $Y$ positive)	$X^p$ where $p < 1$ (e.g., $\ln(X), \sqrt{X}$ , etc.)	$Y^q$ where $q > 1$ (e.g. $Y^2$ and $Y^3$ etc)
Q3 (Both $X$ and $Y$ negative)	$X^p$ where $p < 1$ (e.g. $\ln(X), \sqrt{X}$ , etc.)	$Y^q$ where $q < 1$ (e.g. $\ln(Y), \sqrt{Y}$ , etc.)
Q4 ( $X$ positive and $Y$ negative)	$X^p$ where $p > 1$ (e.g. $X^2, X^3$ , etc.)	$Y^q$ where $q < 1$ (e.g. $\ln(Y), \sqrt{Y}$ , etc.)

In Figure 10.9, the scatter plot is from Quadrant 2 of the Tukey's Bulging Rule and we transformed  $X$  to  $\ln(X)$  which removed the pattern in the residual plot.

### SUMMARY

1. Multiple linear regression model is used to find existence of association relationship between a dependent variable and more than one independent variables.
2. In MLR, the regression coefficients are called partial regression coefficients, since it measures the change in the value of dependent variable for every one unit change in the value of independent variable when all other independent variables in the model are kept constant (*Ceteris Paribus*).
3. When a new variable is added to an MLR model, the increase in  $R^2$  is given by the square of the part-correlation (semi-partial correlation) between the dependent variable and the newly added variable.

4. While building an MLR model, categorical variables with  $n$  categories should be replaced with  $(n - 1)$  dummy (binary) variables to avoid model mis-specification.
5. Apart from checking for normality, heteroscedasticity, every MLR model should be checked for presence of multi-collinearity, since multi-collinearity can destabilize the MLR model.
6. If the data is time-series, then there could be auto-correlation which can result in adding a variable which is statistically not significant. The presence of auto-correlation can result in inflating the  $t$ -statistic value.
7. Variable selection is an important step in MLR model building. Several strategies such as forward, backward, and stepwise variable selection are used in building the model. Partial  $F$ -test is one of the frequently used approached for variable selection.

## Case Study

### Pricing of Players in the Indian Premier League<sup>1</sup>

The year 2008 was a game changer for cricket as a sport. The Indian Premier League (IPL), the professional cricket league tournament initiated by the Board of Control for Cricket in India (BCCI), changed cricket forever. It significantly increased the entertainment value of the game as well as the players' remuneration. Three of the nine IPL teams appeared in the list of 50 highest paying sports teams in the world in a survey conducted by sportingintelligence.com<sup>2</sup> in May 2012 across 278 teams in 14 professional leagues. The salaries paid by some of the IPL teams were far better than what was paid by many popular football clubs in Europe and some of the basketball clubs of the National Basketball Association (NBA), USA. The prices of the IPL players ranged from USD 20,000 to more than USD 2 million for a tournament that was played over 7 weeks.

- The right price for a player and the factors that influenced the pricing puzzled many sports analysts. For example, Suresh Raina and Lasith Malinga had scored more runs and taken more wickets, respectively, in the IPL history, but were not paid as much as some of the other players in the league. Prices were inflated if more than one team was interested in a player. For example, during the auction held on February 4, 2012, Ravindra Jadeja was sold to the Chennai Super Kings (CSK) for more than USD 2 million. Both CSK and Deccan Chargers (DC) bid for Ravindra Jadeja. As of May 14, 2012, Ravindra Jadeja had scored 174 runs at an average of 17.40 and had taken 10 wickets in 14 innings. Whether this performance deserved more than USD 2-million paycheck was debatable. Kolkata Knight Riders purchased Gautam Gambhir for a record USD 2.4 million in Season 4, which was the highest auction price until 2012.

The IPL teams had several restrictions on how much money they could spend on acquiring their players, and it was important that they spend their money wisely to put together a successful team, after gauging the true value of a player.

### Indian Premier League

The Indian Premier League was a professional league for Twenty20 (T20) cricket championships (see **Exhibit 1**) that was started in 2008 in India. The IPL was initiated by the BCCI with eight franchises

<sup>1</sup> ©Indian Institute of Management Bangalore, This case study is authored by U Dinesh Kumar and Kshitiz Ranjan and is distributed through Harvard Business Publishing. This case is not intended to serve as an endorsement, source of primary data, or to show effective or inefficient handling of decision or business processes. Reproduced with permission of IIM Bangalore.

<sup>2</sup> Source: [http://espn.go.com/espn/story/\\_/id/7850531/espn-magazine-sportingintelligence-global-salary-survey-espn-magazine](http://espn.go.com/espn/story/_/id/7850531/espn-magazine-sportingintelligence-global-salary-survey-espn-magazine)

**Continued...**

comprising players from across the world. The first IPL auction was held in 2008 for ownership of the teams for 10 years, with a base price of USD 50 million. In 2012, the IPL consisted of nine teams, and the brand value of the IPL based on the four seasons starting from 2008 was estimated at USD 3.67 billion. Sportingintelligence.com rated the IPL as the second highest paid league after the NBA on a pro-rata basis. In 2009, *Forbes* magazine reported that the IPL was the fastest appreciating sports business in the world.<sup>3</sup> The broadcasting rights of the IPL matches for 10 years were sold to Sony Entertainment Network for USD 1 billion.<sup>4</sup> Delhi Lease & Finance (DLF) Limited, the Indian real estate giant, paid USD 40 million to be the title sponsor of the IPL for five years.<sup>5</sup> The IPL had the backing of prominent Indian industrialists and celebrities adding glamour to the tournament. The owners of the IPL teams ranged from well-known industrialists such as Mukesh Ambani, Chairman of Reliance Industries, and Vijay Mallya, Chairman of the UB Group, to Bollywood actors such as Shah Rukh Khan, Preity Zinta, and Shilpa Shetty. The IPL franchises, their owners, and the corresponding value are shown in **Exhibit 2**. Within a short time, IPL has seen several controversies. The most prominent one was the removal of the founding chairman of the IPL, Lalit Kumar Modi, from his post based on corruption charges. One of the franchises, the Kochi Kerala Tuskers, which played in the 2011 IPL matches, was terminated by the BCCI for breaching the contract.

In the tournament, all the teams played against one another in several matches, at home and away. The teams received 2 points for a victory and 1 point for no result. The top four teams at the league stage played in the semi-finals, and the winning teams from the semi-finals played in the finals. The winner of the finals was declared as the winner of the league. The top three teams from the IPL would also get a chance to compete in the Champions League Twenty20 tournament with the league champions from other countries.

### IPL Player Auctions

The franchises acquired players through an English auction that was conducted every year. However, there were several rules imposed by the IPL. For example, only international players and popular Indian players were auctioned. A base player fee at which the bidding would start was fixed for each player. In the 2008 auction, the franchises were allowed to spend a maximum of USD 5 million in the auction. The players were graded into different baskets<sup>6</sup> based on base price band, specialty, and availability. The players were paid on a pro-rata basis if they were not available for the entire tournament owing to other commitments. The franchises were allowed to acquire players who were not part of the auction; for instance, many Indian players who never played for India were not part of the auction. The money spent on players acquired through the non-auction route was not part of the USD 5 million auction cap. The average weekly salary paid by different franchises based on 2012 survey ranged between USD 56,000 and USD 80,000 (**Exhibit 3**).

<sup>3</sup> Source: <http://www.forbes.com/2009/08/27/cricket-india-ipl-business-sports-ipl.html>

<sup>4</sup> Source: <http://www.ft.com/intl/cms/s/0/a3f80616-0acf-11df-b35f-00144feabdc0.html#axzz1wH4UgkIm>

<sup>5</sup> Source: <http://www.business-standard.com/india/news/dlf-wins-title-sponsorship-rights-for-ipl/313719/>

<sup>6</sup> Source: Singh, Sanjeet, Shaurya Gupta, and Vibhor Gupta, Dynamic bidding strategy for players in IPL auctions, *International Journal of Sports Science and Engineering*, 2011, 5(1), 3–16.

**Continued...**

The auction itself was conducted basket-wise, with breaks between baskets for the franchises to assess their bidding strategy.<sup>7</sup> In the 2008 auctions, five Indian players were nominated as Icon players: Rahul Dravid (Bangalore Royal Challengers), Sourav Ganguly (Kolkata Knight Riders), Virender Sehwag (Delhi Daredevils), Yuvraj Singh (Kings XI Punjab), and Sachin Tendulkar (Mumbai Indians). Each Icon player received 15% more money than the franchise's highest-bid player in the bidding. Further, each franchise had to select at least four players who were under 22 years of age, and the franchise was allotted a catchment area for acquiring local talent. The franchises could select a maximum of 10 non-Indian players; however, only four non-Indian players could be part of the playing 11.

The teams could also buy Indian players outside these annual auctions; the price paid to these players who were not part of the annual auction was not revealed. The teams also retained a few of their players; again, the actual amount paid by a team to retain a player was not publically available.

### The Pricing Challenge

The price of the players in any sports event is driven by many factors. Not all the factors that drove the price of a player are directly related to their performance on the field. For example, David Beckham, who played for LA Galaxy, took home 31.5 million euros in 2012, mainly because of advertisements and endorsement contracts totalling 26 million euros.<sup>8</sup> Beckham's performance on the football field, in terms of the number of goals scored or the number of assists, in 2010–2011 was nothing to write about compared to the on-field performance of Barcelona's Lionel Messi or Real Madrid's Cristiano Ronaldo.

The price was also driven by the marketability of the player; anecdotal evidence suggested that whenever Sachin Tendulkar got out, the television rating points (TRPs) fell sharply since the fans were switching off the television sets. The British Broadcasting Corporation (BBC) quoted that when Sachin Tendulkar went out to bat, people switched on their television sets and switched off their lives.<sup>9</sup> For several years, Tendulkar was one of the biggest brand ambassadors in India. We interviewed former players, administrators, and domain experts to understand their views on the factors that influenced a player's performance.

Sanjay Manjrekar, who played for India in 37 Test matches and 74 One-Day International matches over a period of 10 years, said:

Recent form and the ability to play T20 matches are the major factors that influence a player's price. Franchise owners are very conscious of the T20 ability of the players; very good Test match players may not be selected if their ability in the T20 format is not good. Players who are not worth investing in are left out, especially aging players. I am impressed by how quickly teams have learnt to pick the right players. Some players though are highly overpaid.

<sup>7</sup> Ibid.

<sup>8</sup> Source: <http://www.dawn.com/2012/03/20/messi-worlds-highest-paid-footballer-magazine.html>

<sup>9</sup> Source: [http://en.wikiquote.org/wiki/Sachin\\_Tendulkar](http://en.wikiquote.org/wiki/Sachin_Tendulkar)

**Continued...**

Sujith Somasunder, who represented India in One-Day International matches and was the coach of the Kerala Ranji team, said:

Teams closely watch players' performances in limited over matches. Strike rate is a major factor that influences a player's pricing. Players who play cameo innings and finishers are preferred by the teams.

Rajesh, sports analyst and Stats Editor with *ESPN Cric info*, said:

For a batsman, the ability to score quickly and the flexibility to play anywhere in the line-up are important factors. For bowlers, strike rate and economy rates are important. Fielding skills are a bonus. Players like A.B. de Villiers of South Africa are ideal for this format. He is a top class aggressive batsman who can bat at any position as well as a good fielder who can also keep wickets, and he can captain the side if necessary. A few of the players have fetched far more than they are worth, especially a few Indian players.

The performance of the players could be measured through several metrics. Notably, although the IPL followed the Twenty20 format of the game (refer to **Exhibit 1** for different formats of cricket), it was possible that the performance of the players in the other formats of the game such as Test and One-Day matches could influence player pricing. A few players had excellent records in Test matches, but their records in Twenty20 matches were not very impressive. For example, players such as Ricky Ponting of Australia and V. V. S. Laxman of India had performed very well in Test matches, but did not perform well in T20 matches. The metrics that could be used for measuring player's performance are listed below along with their description. The performance of 130 players played in at least one season of the IPL (2008–2011) measured through various performance metrics are provided in **Table 10.34**. The data description is provided in Table 10.34. Primary objective of this case is to find whether predictive analytics tools such as multiple-linear regression can be used for developing hedonic pricing models for IPL players? In 2013, 37 players were bought through auction by various IPL teams, the performance of 24 of these players along with their base price and sold price is provided in Excel Spreadsheet.<sup>10</sup>

**TABLE 10.34** Data code description

Data Code	Description
AGE	Age of the player at the time of auction classified into three categories: Category 1 ( $L_{25}$ ) means the player is less than 25 years old, category 2 ( $B_{25-35}$ ) means that the age is between 25 and 35 years and category 3 ( $A_{35}$ ) means that the age is more than 35.
RUNS-S	Number of runs scored by a player
RUNS-C	Number of runs conceded by a player

(Continued)

<sup>10</sup> The accompanying excel spreadsheet "IMB381IPL2013.xls".

**Continued...**

**TABLE 10.34** Data code description—Continued

Data Code	Description
HS	Highest score by a batsman in IPL
AVE-B	Average runs scored by a batsman in IPL
AVE-BL	Bowling average (Number of runs conceded/number of wickets taken) in IPL.
SR-B	Batting strike rate (ratio of the number of runs scored to the number of balls faced) in IPL
SR-BL	Bowling strike rate (ratio of the number of balls bowled to the number of wickets taken) in IPL
SIXERS	Number of six runs scored by a player in IPL
WKTS	Number of wickets taken by a player in IPL
ECON	Economy rate of a bowler (number of runs conceded by the bowler per over) in IPL
CAPTAINCY EXP	Captained either an T20 team or a national team
ODI-SR-B	Batting strike rate in One Day Internationals
ODI-SR-BL	Bowling strike rate in One Day Internationals
ODI-RUNS-S	Runs scored in One Day Internationals
ODI-WKTS	Wickets taken in One Day Internationals
T-RUNS-S	Runs scored in Test matches
T-WKTS	Wickets taken in Test matches
PLAYER-SKILL	Player's primary skill (batsman, bowler, or all-rounder)
COUNTRY	Country of origin of the player (AUS: Australia; IND: India; PAK: Pakistan; SA: South Africa; SL: Sri Lanka; NZ: New Zealand; WI: West Indies; OTH: Other countries)
YEAR-A	Year of Auction in IPL
IPL TEAM	Team(s) for which the player had played in the IPL (CSK: Chennai Super Kings; DC: Deccan Chargers; DD: Delhi Daredevils; KXJ: Kings XI Punjab; KKR: Kolkata Knight Riders; MI: Mumbai Indians; PWI: Pune Warriors India; RR: Rajasthan Royals; RCB: Royal Challengers Bangalore). A + sign was used to indicate that the player had played for more than one team. For example, CSK+ would mean that the player had played for CSK as well as for one or more other teams.

## Exhibit 1 Cricket

Cricket is a team sport where a team consisted of 11 players; a team would include a mix of batsmen, bowlers, a wicket keeper, and all-rounders (players with more than one skill). Cricket is the second most popular team sport after soccer, mainly owing to its popularity in the Indian sub-continent.<sup>11</sup> In a cricket match, one team batted first, trying to score as many runs as possible, while the opposing team tried to dismiss the batsmen and limit the number of runs scored. The cricket pitch was usually located in the middle of the playing ground. The cricket pitch consisted of wickets

<sup>11</sup> Source: <http://www.mostpopulairsports.net/>

**Continued...**

(comprising three stumps with bails on the top) on either end of the pitch. The distance between the wickets was 22 yards. Cricket matches were divided into periods called innings; during each innings, one team batted, trying to score as many runs as possible, while the other team bowled and tried to restrict the runs scored. There were many versions of the sport such as Test, One Day International (ODI), and Twenty20 (T20) matches.

In a Test match, each team could bat twice (two innings), and the maximum duration of a Test match was 5 days. An innings got over when 10 of the 11 players of one team got out, or a team could declare the innings to be over before 10 of their players were dismissed, especially if the team thought that they had scored sufficient runs to win the match. Each day, 90 overs consisting of 6 deliveries per over had to be bowled. The team that dismissed the opposition team twice within 5 days and scored more runs than the opposition would be the winner. If a team could not dismiss the opposition team twice within 5 days, the match would be declared a draw.

In a one-day match, each team played for 50 overs, and the team that scored more runs was the winner. If both teams scored the same number of runs, the match would be declared a draw.

In a Twenty20 match, each team played for 20 overs, and the team that scored more runs was the winner. If both teams scored the same number of runs, the match would be declared a draw.

In limited over matches, the winning target for the team batting second was revised in rain interrupted matches using Duckworth and Lewis procedure<sup>12</sup>.

There were several other rules in cricket. For example, in a Twenty20 match, for the first six overs, only two players could field outside the 30-yard circle (from the pitch). Comprehensive information about cricket is available at the following website:

<http://www.cs.purdue.edu/homes/hosking/cricket/explanation.htm>

**EXHIBIT 2** IPL teams and their auction value

Team	Value in USD (million)
Mumbai Indians	111.9
Royal Challengers Bangalore	111.6
Deccan Chargers	107.01
Chennai Super Kings	91
Kolkata Knight Riders	87
Delhi Daredevils	70.4
Kings XI Punjab	66
Rajasthan Royals	58

<sup>12</sup> FC Duckworth and AJ Lewis, "A successful operational research intervention in one-day cricket", Journal of Operational Research Society, Vol. 55, 749–759, 2004.

Continued...

**EXHIBIT 3** Average weekly pay of IPL teams<sup>13</sup>

Team	Average Weekly Pay (in USD)
Kolkata Knight Riders	80,134
Mumbai Indians	79,643
Pune Warriors India	73,839
Royal Challengers Bangalore	71,786
Delhi Daredevils	69,328
Chennai Super Kings	68,373
Deccan Chargers	63,170
Rajasthan Royals	62,455
Kings XI Punjab	56,652

**CASE QUESTIONS**

(use data set IMB381IPL2013.Xls)

1. Develop a simple linear regression model using sold price as response variable and the batting strike rate as explanatory variable. Comment on the model, is there an evidence to suggest that the batting strike rate has a statistically significant relationship with sold price.
2. Develop a multiple linear regression model between sold price and the batting strike rate and number of sixers. Compare this model with the model in question 1. What conclusions can you reach based on both these models?
3. Using model 1, test whether the sold price increases by at least \$1000 for every unit increase in the batting strike rate at significance level of 0.05.
4. Develop a regression model with ideal number of predictors in the model using Mallows's  $C_p$ .

**Suggested Answers to the Case Questions**

1. Develop a simple linear regression model using sold price as response variable and the batting strike rate as explanatory variable. Comment on the model, is there an evidence to suggest that the batting strike rate has a statistically significant relationship with sold price.

**Answer:** The regression model between  $Y = \text{Sold Price}$  and  $X_1 = SR-B$  (batting strike rate) is given in Table 10.35.

<sup>13</sup> Source: [http://espn.go.com/espn/story/\\_/id/7850531/espn-magazine-sportingintelligence-global-salary-survey-espn-magazine](http://espn.go.com/espn/story/_/id/7850531/espn-magazine-sportingintelligence-global-salary-survey-espn-magazine)

Continued...

**TABLE 10.35** Coefficients<sup>a</sup>

Model	Unstandardized Coefficients		Standardized Coefficients	<i>t</i>	Sig.
	<i>B</i>	Std. Error	Beta		
1	(Constant)	289521.427	114770.006	2.523	.013
	SR-B	2086.394	983.643		

<sup>a</sup>Dependent Variable: Sold Price(US\$).

The regression model is

$$Y = 289521.427 + 2086.394 SR-B$$

That is, for every one-unit increase in batting strike rate, the sold price of player increases by \$2086.394. Based on the *p*-value (=0.036) corresponding to the *t*-test, we may conclude that there is a statistically significant relationship between the batting strike rate and the sold price. However, we have to test for the normality and homoscedasticity of the residuals before making this conclusion.

2. Develop a multiple linear regression model between sold price as dependent variable and the batting strike rate (*SR-B*) and number of sixers as independent variables. Compare this model with model in question 1. What conclusions can you reach based on both these models?

**Answer:** The model output, after including sixers in addition to batting strike rate, is shown in Table 10.36.

**TABLE 10.36** Coefficients<sup>a</sup>

Model	Unstandardized Coefficients		Standardized Coefficients	<i>t</i>	Sig.
	<i>B</i>	Std. Error	Beta		
1	(Constant)	395337.539	106613.879	3.708	.000
	SR-B	-102.524	991.030		
	SIXERS	7758.805	1494.312		

<sup>a</sup>Dependent Variable: Sold Price(US\$).

The corresponding model is given by

$$Y = 395337.539 - 102.524 SR-B + 7758.805 SIXERS$$

In model 2 ( $Y = \beta_0 + \beta_1 SR-B + \beta_2 SIXERS$ ), the coefficient for batting striking rate (*SR-B*) is negative which is opposite of what we would expect. Also, in model 1 with just *SR-B* as explanatory variable, the coefficient of *SR-B* is 2086.394. After including the variable *SIXERS*, we see a significant change the coefficient of *SR-B* (from 2086.394 to -102.524). This may be due to multi-collinearity. Whenever there is multi-collinearity, we can expect a huge change the regression coefficient when a new variable is added. Also the regression coefficient may have opposite sign as shown in this case. Also note that the *p*-value for the variable *SR-B* is very high in Table 10.36.

**Continued...**

3. Using model 1, test whether the sold price increases by at least \$1000 for every unit increase in the batting strike rate at significance level of 0.05.

**Answer:** Assume that  $\beta_{SR-B}$  is the coefficient for batting strike rate. This is equivalent to testing the hypothesis testing where

$$\begin{aligned} H_0: \beta_{SR-B} &\leq 1000 \\ H_1: \beta_{SR-B} &> 1000 \end{aligned}$$

The corresponding  $t$ -statistic is given by

$$t = \left( \frac{2086.394 - 1000}{983.643} \right) = 1.104$$

The corresponding  $p$ -value is 0.1357.

Since the  $p$ -value is greater than 0.05, we retain the null hypothesis and conclude that the sold price is unlikely to increase by more than \$1000 for every one-unit increase in the batting strike rate at  $\alpha = 0.05$ .

4. Develop a regression model with ideal number of predictors in the model using Mallows's  $C_p$ .

**Answer:** We have to develop a stepwise regression model while calculating  $C_p$  at each step. Tables 10.37 and 10.38 show output from SAS Enterprise Miner where the value of  $C_p = 7.0252$  after including 7 variables and the corresponding  $R$ -square is 0.5969. Since  $C_p = 7.0252$  and  $p = 8$  (including constant), we choose the model in Table 10.38 as the best model. For any other model, the difference between  $C_p$  and  $p$  will higher than the model in Table 10.38.

In Table 10.38,  $BOW \times WK - O$  is the interaction between bowler and wickets taken in one international matches.

**TABLE 10.37** IPL pricing regression summary

Source	df	Analysis of Variance			
		Sum of Squares	Mean Square	F-Value	Pr > F
Model	7	1.274364E13	1.82052E12	25.81	< 0.0001
Error	122	8.604856E12	70531604276		
Corrected Total	129	2.13485E13			
R-Square	0.5969	$C_p$		7.0252	

**TABLE 10.38** Sold price vs player performance

Variable	Parameter Estimate	Standard Error	Type II SS	F-Value	Pr > F
Intercept	-50754	52320	66372008548	0.94	0.3339
$L_{25}$	268664	96885	5.423618E11	7.69	0.0064

**- Continued ...**

**TABLE 10.38** Sold price vs player performance—Continued

Variable	Parameter Estimate	Standard Error	Type II SS	F-Value	Pr > F
BOW*WK-0	-937.88374	385.23153	4.180595E11	5.93	0.0164
ODI-WKTS	1233.92565	342.63784	9.147262E11	12.97	0.0005
INDIA	223438	52740	1.265937E12	17.95	<.0001
SIXERS	5510.23860	1065.67323	1.885717E12	26.74	<.0001
Year_Dummy	138482	54030	4.633361E11	6.57	0.0116
Base Price(US\$)	1.39414	0.16293	5.164098E12	73.22	<.0001

## MULTIPLE CHOICE QUESTIONS

7. When a stepwise regression model is developed, the first variable that is added is
  - (a) The variable with highest variance.
  - (b) The variable that has the least variance.
  - (c) The variable that has highest correlation with the dependent variable.
  - (d) The variable with least covariance with the dependent variable.
8. Variance inflation factor is
  - (a) Factor by which the regression coefficient is increase(d)
  - (b) Factor by which the  $t$ -statistic value is inflate(d)
  - (c) Factor by which the  $t$ -statistic is deflated by a factor of  $\sqrt{VIF}$ .
  - (d) Factor by which the  $t$ -statistic value is inflated by a factor of  $\sqrt{VIF}$ .
9. Variable selection in stepwise regression is achieved through:
  - (a) Partial F-test
  - (b)  $F$ -test
  - (c) Correlation
  - (d)  $t$ -test
10. A regression model is developed between salary earned by a graduating MBA student using a sample of 450 students and their undergraduate discipline (where the base category is discipline “arts”). The regression output is shown in Table 10.39.

**TABLE 10.39** Regression coefficients

Model	Unstandardized Coefficients		<i>t</i>	Sig.
	<i>B</i>	Std. Error		
1	(Constant)	198246.40	45690.10	4.338
	Science	39430.00	20020.60	2.121
	Engineering	56940.50	22450.67	2.536
	Commerce	-14250.89	8932.45	1.5954

Which of the following statements are true at 5% significance:

- (a) Students from arts category earn minimum average salary.
- (b) Students from engineering category earn the maximum average salary.
- (c) The average salaries earned by arts and commerce graduates are same.
- (d) Science students earn 39430 more than arts students on average.
11. A regression model is developed for salary of employees of a company using gender (*G*), work experience (*WE*) and the interaction variable  $G \times WE$ .  $G = 1$  is coded as female and  $G = 0$  is male. The corresponding regression equation is shown below (assume that all predictors are significant):

$$Y = 45,490.50 + 3000.900 \times G + 1497.89 \text{ } WE - 990.75 \text{ } G \times WE$$

Which of the following statements are true?

- (a) Average salary of female employees is higher than male employees
- (b) Female employees earn 3000.90 more than male employees
- (c) Increase in salary with work experience for male employees is higher than female employees.
- (d) In the long run, male employees earn more than female employees.
12. Which of the following measures are used for identifying influential observations in the data?
  - (a) Cook's distance
  - (b) Mahalanobis distance
  - (c) Leverage value
  - (d) All of above

13. Transformation of variables will be useful to solve the following problem(s) in MLR:
- Multi-collinearity
  - Outliers
  - Heteroscedasticity
  - None of above
14. Regression model was developed on a time-series data, the value of Durbin–Watson statistic value is 0.2. Then
- There is a significant correlation between the independent variable and dependent variable.
  - There is a positive auto-correlation between errors.
  - There is a negative auto-correlation between errors.
  - There is no auto-correlation.
15. The independent variable that has the highest impact on the dependent variable is given by
- The variable with largest coefficient value.
  - The variable with largest absolute coefficient value.
  - The variable with largest standardized coefficient value.
  - The variable with largest absolute standardized coefficient value.

### EXERCISES

**Data for Questions 1–6:** The dean of a business school has collected data on their recent placement. To attract good students, it is important for the school to ensure that the students are placed with good salary package. The dean of the school believed that the salary earned by a student at placement depended on several variables. The data collected by the dean is listed in Table 10.40.

TABLE 10.40 Data description

S. No.	Variable	Variable Type	Code in SPSS output
1	Salary ( $Y$ )	Numerical	Salary
2	Gender	Categorical	Gender = 1 (Male), 0 (Female)
3	Percentage Marks in SSC	Numerical	Percent_SSC SSC_CBSE
4	Board SSC	Categorical (3 levels)	SSC_ICSE SSC_OTHERS
5	Percentage Marks in HSC	Numerical	Percent_HSC
6	Percentage Marks in Degree	Numerical	Percent_Degree Degree_Arts Degree_Commerce
7	Degree Specialization	Categorical (6 levels)	Degree_CompApp Degree_Engineering Degree_Science Degree_Management

(Continued)

**TABLE 10.40** Data description—Continued

S. No.	Variable	Variable Type	Code in SPSS output
8	Years of Experience	Numerical measured in years	Experience_Yrs
9	Entrance Exam	Categorical	$ENT = 1$ implies took entrance exam $ENT = 0$ implies otherwise
10	Percentage in MBA	Numerical	Percent_MBA
11	Marks in communication	Numerical	Marks_Communication

The first regression model is built using degree of specialization as the explanatory variable.

### MODEL 1

$$Y = \beta_0 + \beta_1 \text{Degree\_Arts} + \beta_2 \text{Degree\_Commerce} + \beta_3 \text{Degree\_CompApp} + \beta_4 \text{Degree\_Engineering} + \beta_5 \text{Degree\_Management}$$

The model 1 SPSS outputs are shown in Tables 10.41– 10.43.

**TABLE 10.41** Model summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	271 <sup>a</sup>	.073	.058	82949.958

<sup>a</sup>Predictors: (Constant), Degree\_Management, Degree\_Arts, Degree\_CompApp, Degree\_Engineering, Degree\_Commerce.

**TABLE 10.42** ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
	Regression	1.662E11	5			
1	Residual		305			
	Total	2.265E12	310			

<sup>a</sup>Dependent Variable: Salary.

**TABLE 10.43** Coefficients<sup>a</sup>

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1	(Constant)	261440.000	16589.992		
	Degree_Arts	-14040.000	31037.032		
	Degree_Commerce	26294.043	18666.192		
	Degree_CompApp	13393.333	23704.925		
	Degree_Engineering	63760.000	22462.955		
	Degree_Management	-9013.437	18137.895		

<sup>a</sup>Dependent Variable: Salary.

- Assuming that the salary package is important for the school, should the dean give more importance to certain degree disciplines while admitting the students to their MBA programme? Support your answers with precise arguments.
- Is there a significant difference between the average salary earned by a student with science degree and commerce degree? Clearly state your arguments.
- The dean of the school believes that the engineering students earn on average at least INR 25,000 more than the science students. Check whether his belief is true at 5% significance level by conducting an appropriate hypothesis tests.

A new variable, which is the interaction between degree discipline engineering and the percentage marks in degree, is added to model 1 and the corresponding output is shown in Table 10.44.

**TABLE 10.44** Coefficients

Model	Unstandardized Coefficients		<i>t</i>	Sig.	VIF
	<i>B</i>	Std. Error			
1	(Constant)	261440.000	16520.960	15.825	0.000
	Degree_Arts	-14040.000	30907.885	-0.454	0.650
	Degree_Commerce	26294.043	18588.520	1.415	0.158
	Degree_CompApp	13393.333	23606.287	0.567	0.571
	Degree_Engineering	336963.387	146632.427	2.298	0.022
	Degree_Management	-9013.437	18062.423	-0.499	0.618
	ENGPERCENT <sup>a</sup>	-5444.138	2357.318	-2.309	0.021

<sup>a</sup>ENGPERCENT is interaction between Degree\_Engineering and Percent\_Degree.

- Interpret the coefficient value for the interaction value *ENGPERCENT* (*Degree\_Engineering*  $\times$  *Percent\_Degree*). Explain possible reason for the salary of engineering students decreasing as the percentage marks in degree increases. Clearly state your arguments.

A stepwise regression is carried out using SPSS and the results of stepwise regression are shown in Tables 10.45 and 10.46.

**TABLE 10.45** Model summary

Model	<i>R</i>	<i>R-Square</i>	Adjusted <i>R-Square</i>	Std. Error of the Estimate
1	0.246 <sup>a</sup>		0.057	82984.946
2				
3				
4				

**TABLE 10.46** Coefficient values

Model	Unstandardized Coefficients		Sig.	Correlations		
	B	Std. Error		Zero-order	Partial	Part
1	(Constant)	131027.092	32059.466	0.000		
	Marks_Communication	2333.254	523.349	0.000	0.246	0.246
2	(Constant)	96461.563	32253.883	0.003		
	Marks_Communication	2441.930	510.130	0.000	0.246	0.263
	GENCOM	689.203	162.261	0.000	0.215	0.235
3	(Constant)	116685.273	32465.888	0.000		
	Marks_Communication	2323.885	504.517	0.000	0.246	0.254
	GENCOM	658.158	160.332	0.000	0.215	0.228
	Degree_Management	-28695.590	9222.060	0.002	-0.196	-0.175
4	(Constant)	116712.754	32228.770	0.000		
	Marks_Communication	2242.984	502.012	0.000	0.246	0.247
	GENCOM	629.520	159.625	0.000	0.215	0.220
	Degree_Management	-22777.435	9494.078	0.017	-0.196	-0.136
	Degree_Engineering	37336.093	15871.087	0.019	0.202	0.133

5. What is the  $R$ -square value at step 2 of the stepwise regression?
6. In Table 10.46, GENCOM is the interaction variable between gender and marks in communication. Which of the following statements is true? Clearly state your arguments.
  - (a) Salary is more sensitive to marks in communication for females than males.
  - (b) Salary is more sensitive to marks in communication for males than females.
  - (c) There is no difference between males and females with respect to marks in communication.
  - (d) Can't say.

**Data for Questions 7–12 (Courtesy: Professor Trilochan Sastry, IIM Bangalore):** An Agro Insurance company wanted to come up with a model and see how the total production of paddy depends on the rainfall. The complication is that the productivity also depends on various factors such as the total acreage under irrigation. The following variables were used to develop the regression models:

PROD	The total production in thousands of tons (dependent variable)
IRR	Total irrigated area in thousands of hectares (independent variable)
NON	Total non-irrigated area in thousands of hectares (independent variable)
RAIN	Total rainfall in millimetres (independent variable)

The SPSS regression model output is given Table 10.47.

**TABLE 10.47** Regression model output

Model	R	R-Square	Adjusted R-Square	Std. Error of the Estimate	Degrees of freedom SSR	Degrees of freedom SSE
Model 1	0.895	0.801	0.787	703.6283	3	44

7. If stepwise regression was used to arrive at Table 10.47, how many variables did SPSS consider? Give reasons.
8. How many observations were included in the regression?
- ANOVA corresponding to the MLR model developed is shown in Table 10.48.

**TABLE 10.48** ANOVA<sup>a</sup>

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	8.762E7				.000 <sup>a</sup>
Residual					
Total					

<sup>a</sup>Dependent Variable: PROD.

9. Fill in the blanks in the ANOVA table (Table 10.48) for Sum of Squares, *df*, Mean Square and *F* value.
10. Regression coefficients for the model developed are shown in Table 10.49. Does the regression exhibit any collinearity? Give reasons.

**TABLE 10.49** Coefficients<sup>a</sup>

Model	Unstandardized Coefficients		Beta	t	Sig.	95.0% Confidence Interval for <i>B</i>		Collinearity Statistics	
	<i>B</i>	Std. Error				Lower Bound	Upper Bound	Tolerance	VIF
1 (Constant)	-929.61	1097.238		-0.847	0.401	-3140.948	1281.728		
IRR	2.281	0.192	0.809	11.906	0.000	1.895	2.667	0.979	1.021
NON	0.622	0.159	0.266	3.906	0.000	0.301	0.943	0.977	1.023
RAIN	2.112	0.670	0.213	3.152	0.003	0.761	3.462	0.992	1.008

<sup>a</sup>Dependent Variable: PROD.

11. The normal and residual plots are given in Figures 10.13 and 10.14. Which assumptions of regression are tested by them, and are they satisfied?  
Selected data from the SPSS output is given in Table 10.50.

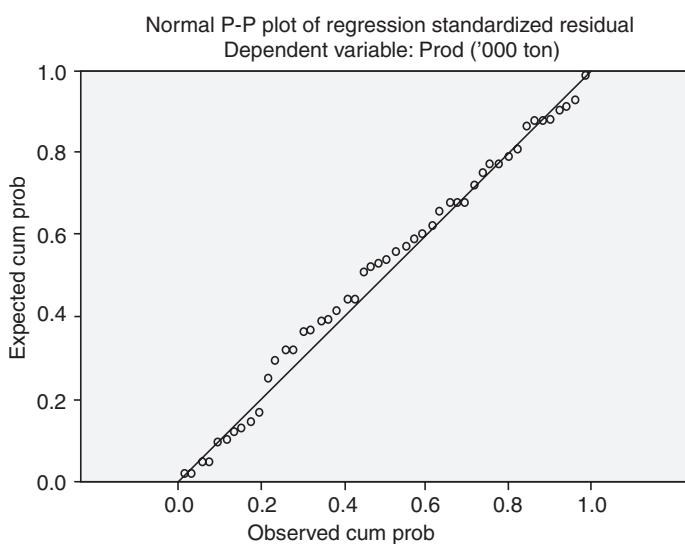
**TABLE 10.50** SPSS output on influential observations for portion of the sample

Year	Prod '000 Ton	ZRE	MAH	COO	LEV	SDFB0	SDFB1	SDFB2	SDFB3
1952–53	2930	0.05093	9.3422	0.0002	0.1988	30.8270	-0.0019	-0.0036	-0.0101
1953–54	3450	-0.28396	14.4868	0.0147	0.3082	-160.9821	0.0090	0.0297	-0.0715
1954–55	4250	1.18262	5.1603	0.0604	0.1098	421.8451	-0.0426	-0.0409	-0.1986
1955–56	3860	0.07316	5.0852	0.0002	0.1082	22.6753	-0.0022	-0.0035	0.0041
1956–57	4370	-0.15086	6.0918	0.0012	0.1296	-17.8381	0.0047	0.0046	-0.0307
1957–58	4710	0.211	1.1081	0.0005	0.0236	8.5996	-0.0064	0.0000	-0.0007
1958–59	5180	1.1752	1.3166	0.0186	0.0280	101.3696	-0.0372	-0.0064	-0.0259
1959–60	4560	-0.21501	1.4856	0.0007	0.0316	1.6727	0.0068	-0.0006	-0.0116

(Continued)

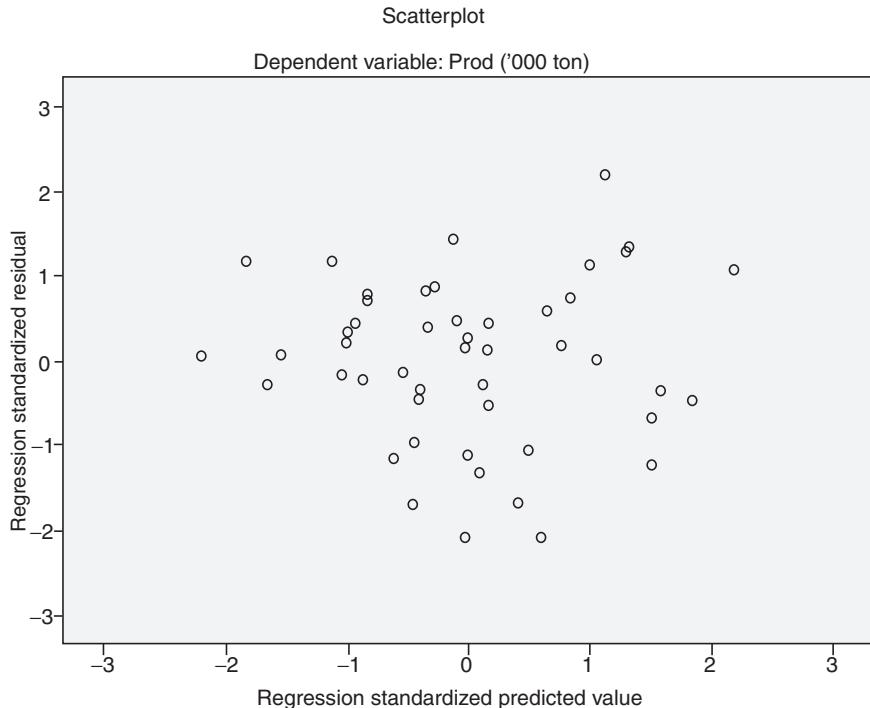
**TABLE 10.50** SPSS output on influential observations for portion of the sample—Continued

Year	Prod '000 Ton	ZRE	MAH	COO	LEV	SDFB0	SDFB1	SDFB2	SDFB3
1960–61	4810	0.32379	1.0865	0.0013	0.0231	11.8070	-0.0098	0.0003	-0.0027
1961–62	4990	0.45626	1.9852	0.0037	0.0422	-2.3511	-0.0147	0.0066	-0.0383
1962–63	5060	-0.14592	2.9114	0.0005	0.0619	27.0517	0.0054	-0.0050	-0.0038
1963–64	5300	0.74572	1.6166	0.0086	0.0344	-32.8552	-0.0246	0.0122	-0.0258
1964–65	6000	0.81488	3.6257	0.0200	0.0771	-198.7153	-0.0290	0.0311	0.0662
1965–66	4260	-1.15514	5.6126	0.0633	0.1194	208.5132	0.0454	-0.0551	0.1539
1966–67	4410	-2.07939	4.1774	0.1496	0.0889	527.7487	0.0528	-0.0634	-0.3594
1967–68	5730	0.40227	2.4797	0.0035	0.0528	-58.1897	-0.0102	0.0133	-0.0209
1968–69	4630	-0.96035	1.0713	0.0110	0.0228	83.6995	0.0228	-0.0170	-0.0143
1969–70	5130	-0.33885	0.8750	0.0012	0.0186	25.1432	0.0071	-0.0056	0.0000

**FIGURE 10.13** Normal P-P plot.

12. Identify observations that are highly influential using information provided in Table 10.50. Explain clearly.

**Data for Questions 13–20:** Box-office collection of 149 Bollywood movies were analysed using the variables described in Table 10.51.



**FIGURE 10.14** Residual plot.

**TABLE 10.51** Data dictionary bollywood box office collection data

S. No.	Variable	Variable Type	Code in SPSS Output
1	Box office Collection (Y)	Numerical (in crores of rupees)	Box Office Collection
2	Release Time	Categorical with 4 levels	Releasing_Time_Festival Season Releasing_Time_Holiday Season Releasing_Time_Long Weekend Releasing_Time_Normal Season
3	Genre	Categorical with 5 levels	Genre_Action (Action) Genre_Drama (Drama) Genre_Romance (Romance) Genre_Comedy (Comedy) Genre_Others (Other-G)
4	Movie Content	Categorical with 3 levels	Masala (Masala) Sequel (Sequel) Others (Other_C)

(Continued)

**TABLE 10.51** Data Dictionary Bollywood Box Office collection data—Continued

S. No.	Variable	Variable Type	Code in SPSS Output
5	Director Category	Categorical with 3 levels	Director_A
			Director_B
			Director_O
6	Lead Actor Category	Categorical with 3 levels	Actor_A
			Actor_B
			Actor_O
7	Music Director Category	Categorical with 3 levels	Music_Dir_CAT A
			Music_Dir_CAT B
			Music_Dir_CAT C
8	Production House Category	Categorical with 3 levels	Prod_House_CAT A
			Prod_House_CAT B
			Prod_House_CAT C
9	Item Song	Binary variable	Item_Song (1 implies that the movie has an item song, 0 otherwise)
10	Budget	Numerical (in crores of rupees)	Budget
11	YouTube Views	Numerical	YouTube-V
12	YouTube Likes	Numerical	YouTube-L
13	YouTube Dislikes	Numerical	YouTube-D
14	Budget More than 35 crores	Categorical	Budget_35_Cr (1 if the budget is more than 35 crores 0 otherwise)

A simple linear regression model was developed between box-office collection and budget. SPSS output of the model is shown in Tables 10.52–10.53 and Figures 10.15–10.16.

### MODEL 1

$$Y \text{ (Box-Office Collection)} = \beta_0 + \beta_1 \times \text{Budget}$$

**TABLE 10.52** Model summary<sup>a</sup>

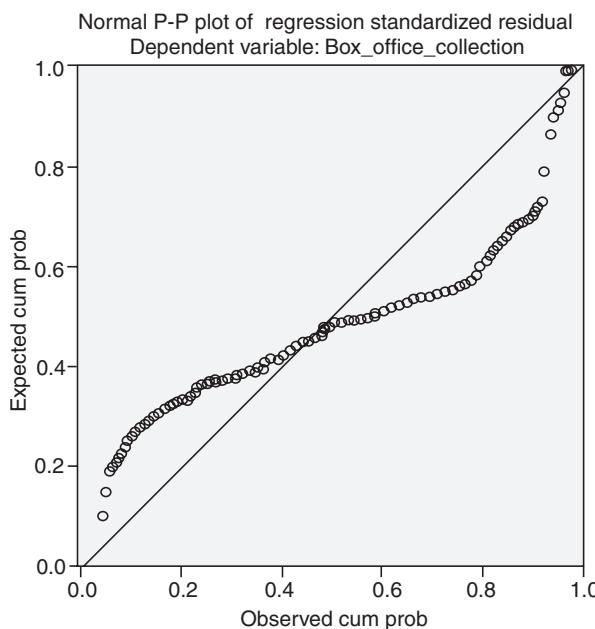
Model	R	R-Square	Adjusted R-Square	Std. Error of the Estimate
1	0.650 <sup>b</sup>			72.02261

<sup>a</sup>Dependent Variable: Box\_Office\_Collection <sup>b</sup>Predictors: (Constant), Budget.

**TABLE 10.53** Coefficients<sup>a</sup>

Model	Unstandardized Coefficients			Standardized Coefficients	
	B	Std. Error	Beta	t	Sig.
1	(Constant) -8.354	8.535		-0.979	0.329
	Budget 2.175	0.210	0.650	10.381	0.000

<sup>a</sup>Dependent Variable: Box\_Office\_Collection.

**FIGURE 10.15** Normal P-P plot for Model 1.

13. Which of the following statements are correct (more than one may be correct)?
- The model explains 42.25% of variation in box-office collection.
  - There are outliers in the model.
  - The residuals do not follow a normal distribution.
  - The model cannot be used since R-square is low.
  - Box office collection increases as the budget increases.
14. Mr Chellappa, CEO of Oho Productions (OP), claims that the regression model in Table (10.56) is incorrect since it has negative constant value. Comment whether Mr Chellappa is correct in his assessment about the model.
- A second model is developed between  $\ln(\text{Box office collection})$  and movie release time:

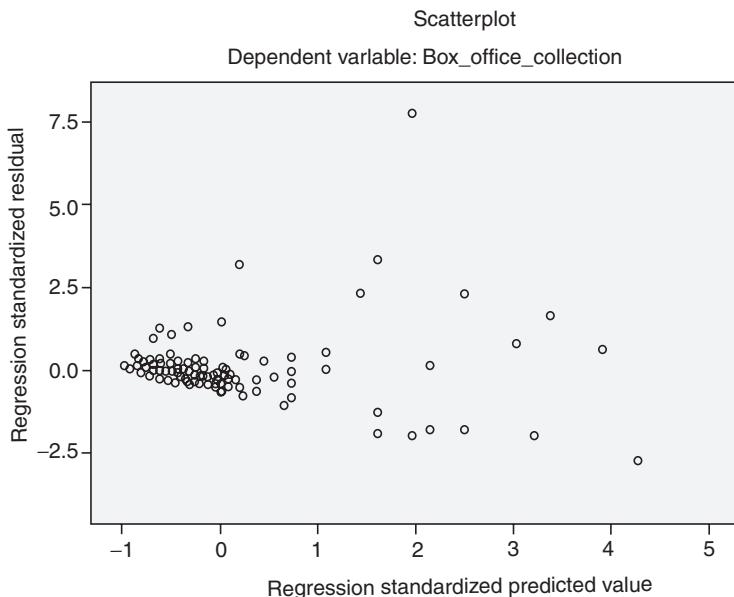


FIGURE 10.16 Residual plot for Model 1.

**MODEL 2**

$$\ln(Y) = \beta_0 + \beta_1 \times \text{Release Time Festival Season} + \beta_2 \times \text{Release Time Long Weekend} \\ + \beta_3 \times \text{Release Time Normal Season} + \varepsilon$$

The regression output for Model 2 is given in Table 10.54.

TABLE 10.54 Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients		t	Sig.
		B	Std. Error	Beta			
2	(Constant)	2.685	0.396			6.776	0.000
	Releasing_Time_Festival_Season	0.727	0.568	0.136		1.278	0.203
	Releasing_Time_Long_Weekend	1.247	0.588	0.221		2.122	0.036
	Releasing_Time_Normal_Season	0.147	0.431	0.041		0.340	0.734

<sup>a</sup>Dependent Variable: Ln(Box Office Collection).

- What is the average difference in the box-office collection when a movie is released during a holiday season (Releasing\_Time\_holiday\_season) versus movies released during normal season (Releasing\_Time\_Normal\_Season)? Use a significance value of 5%.
- Mr Chellappa of Oho productions claims that the movies released during long weekend (Releasing\_Time\_Long\_Weekend) earn at least 5 crores more than the movies released during normal season (Releasing\_Time\_Normal\_Season). Check whether this claim is true (use  $\alpha = 0.05$ ).

A stepwise regression model is developed between  $\ln(\text{Box Office Collection})$  and all the predictor variables listed in Table 10.51. The outputs are shown in Tables 10.55–10.56.

**TABLE 10.55** Model summary

Model	R	R-Square	Adjusted R-Square	Std. Error of the Estimate
1	0.709	0.503	0.499	1.20651
2	0.763	0.581	0.576	1.11050
3	0.787	0.620	0.612	1.06210
4	0.802	0.643	0.633	1.03307
5	0.810			1.01749

**TABLE 10.56** Coefficients in the model (in the order in which it was added to the model)

Model	Unstandardized Coefficients		Beta	t	Correlations		
	B	Std. Error			Zero-Order (Direct)	Partial	Part
Step 6	(Constant)	3.573	.249	14.346			
	Budget_35_Cr	1.523	.207	.443	7.342	.709	.525
	Youtube_VIEWS	1.1710 <sup>-07</sup>	.000	.242	4.426	.538	.348
	Prod_House_CAT A	0.562	.185	.165	3.033	.444	.247
	Music_Dir_CAT C	-0.645	.199	-.177	-3.245	-.483	-.263
	GenreComedy	0.456	.197	.115	2.312	.006	.190
	Director_CAT C	-0.434	.203	-.123	-2.143	-.509	-.177

17. What is the variation in response variable, ln(Box office collection), explained by the model after adding all 6 variables?
18. Which factor has the maximum impact on the box-office collection of a movie? What will be your recommendation to a production house based on the variable that has maximum impact on the box office collection?
19. Compare the regressions in Model 2 (Table 10.54) and Model 3 (Tables 10.55 and 10.56). None of the variables in Model 2 are statistically significant in Model 3. Can we conclude that the variables in Model 2 have no association relationship with Box-Office Collection? Explain clearly.
20. Among the variables in Table 10.56, which variable is not useful for practical application of the model? Clearly state your reasons.

## REFERENCES

1. Beale E M L (1970), "Note on Procedures for Variable Selection in Multiple Regression", *Technometrics*, **12**(4), 909–914
2. Belsley D A , Kuh E, and Welsch R E (1980). "Regression Diagnostics: Identifying Influential Data and Sources of Collinearity", John Wiley and Sons, New York.
3. Bingham C (1977), "Some Identities Useful in the Analysis of Residuals from Linear Regression", Technical Report 300, School of Statistics, University of Minnesota.
4. Box G E P and Cox D R (1964), "An Analysis of Transformations", *Journal of the Royal Statistical Society – Series B (Methodology)*, **26**(2), 211–252

5. Chatterjee S and Hadi A S (1986), "Influential Observations, High Leverage Points and Outliers in Linear Regression", *Statistical Science*, **1**(3), 379–393.
6. Cook R D (1977). "Detection of Influential Observation in Linear Regression", *Technometrics*, **19**(1), 15–18.
7. Durbin J and Watson G S (1950), "Testing for serial correlation in Least Square Regression: I", *Biometrika*, **37**(3/4), 409–428.
8. Durbin J and Watson G S (1951), "Testing for serial correlation in Least Square Regression: II", *Biometrika*, **38**(1/2), 159–177.
9. Durbin J and Watson G S (1971), "Testing for serial correlation in Least Square Regression: III", *Biometrika*, **58**(1), 1–19.
10. Gujarati D N and Sangeetha (2010), "Basic Econometrics", Tata McGraw Hill Education Private Limited, New Delhi
11. Halinski R S and Feldt L S (1970), "The Selection of Variables in Multiple Regression Analysis", *Journal of Educational Measurement*, **7**(3), 151–157.
12. Mahalanobis P C (1936), "On the Generalized Distance in Statistics", *Journal of the Asiatic Society of Bengal*, **2**(1), 49–55.
13. Mallows C L (1973), "Some Comments on Cp", *Technometrics*, **15**(4), 661–675.
14. Rousseeuw P J and Zomeren B C (1990), "Unmasking Multi-variate Outliers and leverage Points", *Journal of the American Statistical Association*, **85**(411), 633–639
15. Ryan T P (2009), "Modern Regression Methods, 2<sup>nd</sup> Edition", John Wiley and Sons, Hoboken, New Jersey.
16. Thompson M L (1978), "Selection of Variables in Multiple Regression: Part I. A Review and Evaluation", *International Statistical Review*, **46**(1), 1–19.
17. Warren R, Smith R E, and Cybenko A K (2011), "Use of Mahalanobis distance for Detecting Outliers and Outlier Clusters in Markedly Non-Normal Data: A Vehicular Traffic Example", *Air Force Research Laboratory Report AFRL-RH-WP-TR-2011*, June 2011.

# 11

# Logistic Regression

“Science is the systemic classification of experience.”

— George Hendry Lewes

## LEARNING OBJECTIVES

- LO 11-1** Understand classification problems and the techniques that are used for solving them.
- LO 11-2** Understand basic concepts in logistic regression and how logistic regression model is developed.
- LO 11-3** Understand the estimation of logistic regression coefficients and interpretation.
- LO 11-4** Learn concepts such as sensitivity, specificity, classification matrix, and classification plot.
- LO 11-5** Learn concepts such as receiver operating characteristic curve (ROC curve), area under ROC curve (AUC), gain chart, and lift chart and its application in model selection.
- LO 11-6** Understand logistic regression model deployment using credit rating example.
- LO 11-7** Learn applications of logistic regression model across several industries.

## ESSENCE OF LOGISTIC REGRESSION

Logistic regression (LR) is a statistical technique for finding existence of a relationship between a qualitative (discrete) dependent variable (or outcome variable) and several independent variables (aka explanatory variables or predictors). In LR, the primary objective is to find the conditional probability of occurrence of an event (class probability) given the values of several independent variables. LR is one of the most popular tools used for solving classification problems.

## 11.1 | INTRODUCTION – CLASSIFICATION PROBLEMS

Classification problems are an important category of problems in analytics in which the response variable ( $Y$ ) takes a discrete value. In classification problems, the primary objective is to predict the class of a customer (or class probability) based on the values of explanatory variables or predictors. Few examples of classification problems are listed below:

1. A bank may like to classify their customers based on risk such as low-, medium- and high-risk customers under loan portfolio. Here the response variable  $Y$  takes 3 values (e.g.,  $Y = 1$  for low risk,  $Y = 2$  for medium risk and  $Y = 3$  for high risk).

2. An organization may like to predict the customers who are likely to churn (here  $Y$  takes two values,  $Y = 1$  for churn and  $Y = 0$  for do not churn).
3. Health service providers based on diagnostic tests may classify the patients as positive, that is presence of a disease ( $Y = 1$ ) or negative, that is absence of a disease ( $Y = 0$ ).
4. Customers who are likely to respond to a marketing campaign through phone calls/emails ( $Y = 1$  will respond to the campaign;  $Y = 0$  will not respond).
5. Human Resource Department of a firm may try to predict whether an applicant would accept the job offer (two categories: accept and do not accept).
6. Movie production houses may like to predict whether a movie will be a hit or not at the box office.
7. Predict outcome of any sporting event, for example, in case of football the outcome will be Win, Draw or Loss.
8. Many organizations such as banks, e-commerce and insurance companies have to deal with fraudulent transactions. They may like to predict whether a transaction is fraud or not.
9. Few companies manipulate the accounts, so the policy makers (or regulators) may like to predict whether the companies are manipulating the accounts or not.
10. Sentiment about a product or service in social media can be classified as positive, negative, or neutral, which enables an organization to understand sentiments about their product/service. Organizations may like to understand the reasons for negative sentiment if exists and take corrective actions.

The above are few examples of classification problems. In general, any experiment that has binary or multinomial outcomes (classes) is called as a classification problem. There are several techniques used for solving classification problems such as logistic regression, classification trees, discriminant analysis, neural networks, and support vector machines. In this chapter, we will discuss binomial logistic regression models.

## 11.2 | INTRODUCTION TO BINARY LOGISTIC REGRESSION

Logistic regression is a statistical model in which the response variable takes a discrete value and the explanatory variables can either be continuous or discrete. Logistic regression is one of the supervised learning algorithms. In this chapter, we will be discussing binary logistic regression in which the response variable takes only two values. For example, assume that the value of  $Y$  is either 1 (conventionally known as positive outcome) or 0 (conventionally known as negative outcome). When there are more than two values of  $Y$ , then multinomial logistic regression model is used. The binary logistic regression model is given by

$$P(Y=1) = \frac{e^Z}{1+e^Z} \quad (11.1)$$

where

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m \quad (11.2)$$

Here  $X_1, X_2, \dots, X_m$  are the independent variables. The right-hand side of Eq. (11.1) is a logistic function (and thus the name logistic regression). One of the objectives of classification problems is to

predict the class probability, that is the probability that an observation will belong to a particular class. Equation (11.1) gives the class probability of an observation belonging to class labelled as 1, that is  $P(Y = 1)$ . Logistic function is a probability function, and has an S-shaped curve as shown in Figure 11.1.

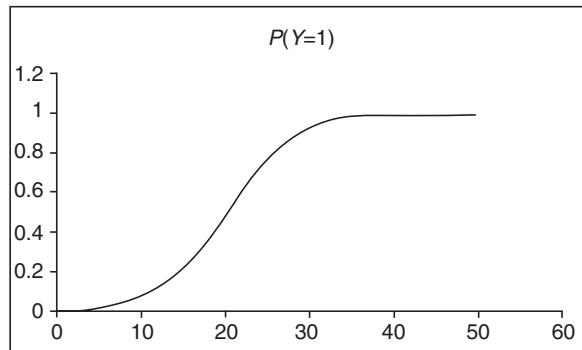


FIGURE 11.1 Logistic function.

The logistic regression function defined in Eq. (11.1) can be transformed as follows:

$$\frac{P(Y = 1)}{1 - P(Y = 1)} = e^z \quad (11.3)$$

Equation (11.3) can be written as

$$\ln\left(\frac{P(Y = 1)}{1 - P(Y = 1)}\right) = Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m \quad (11.4)$$

Now  $\frac{P(Y = 1)}{1 - P(Y = 1)}$  is odds. Thus the left-hand side of Eq. (11.4) is log-natural of odds. Equation (11.4) is known as logit (**logistic probability unit**) function. The left-hand side of logit function is a continuous function and the right-hand side is a linear function. Logit function is similar to a multiple linear regression model. Such models are called *generalized linear models* (GLM), in GLM the errors do not follow normal distribution and there exists a transformation function of the outcome variable that takes a linear function. For example, consider a regression equation (11.5) in which the response variable  $Y$  takes only two values (0 or 1):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i \quad (11.5)$$

The error ( $\varepsilon_i$ ) is given by

$$\varepsilon_i = \begin{cases} 1 - \beta_0 - \beta_1 X_{1i} & \text{for } Y_i = 1 \\ 0 - \beta_0 - \beta_1 X_{1i} & \text{for } Y_i = 0 \end{cases} \quad (11.6)$$

Thus, for a given value of  $X_{1i}$ , the error can take only two values as given in Eq. (11.6) and thus will not follow normal distribution. The logit function [Eq. (11.4)] is known as the *link function*.

### 11.3 | ESTIMATION OF PARAMETERS IN LOGISTIC REGRESSION

One of the major assumptions of multiple linear regression model is that the residuals follow a normal distribution (or approximate normal distribution). However, the residuals in logistic regression will not follow normal distribution [Eq. (11.6)] and thus we cannot use method of ordinary least squares (OLS) to estimate the regression parameters. Also, the errors in LR are inherently heteroscedastic (DeMaris, 1995). Regression parameters in the case of logistic regression are estimated using Maximum Likelihood Estimator (MLE). In binary logistic regression, the response variable  $Y$  takes only two values ( $Y = 0$  and 1). Let

$$P(Y=1|Z=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_mX_m)=\pi(Z)=\frac{e^Z}{(1+e^Z)} \quad (11.7)$$

The probability (likelihood) function of binary logistic regression for specific observation  $Y_i$  ( $Y_i=0$  or 1) is given by

$$P(Y_i)=\pi(Z)^{Y_i}(1-\pi(Z))^{1-Y_i} \quad (11.8)$$

Assume that the data set has  $n$  observations,  $Y_1, Y_2, \dots, Y_n$ . The likelihood function, which is a joint probability,  $L(Y_1, Y_2, \dots, Y_n)$  for a specific  $Z_i (= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_m X_{mi})$  is given by

$$L=P(Y_1,Y_2,...,Y_n)=\prod_{i=1}^n \pi(Z_i)^{Y_i}[1-\pi(Z_i)]^{1-Y_i} \quad (11.9)$$

The log-likelihood function is given by

$$\ln(L)=LL=\sum_{i=1}^n Y_i \ln[\pi(Z_i)]+\sum_{i=1}^n (1-Y_i)[\ln(1-\pi(Z_i))] \quad (11.10)$$

For mathematical simplicity, assume that  $Z_i = \beta_0 + \beta_1 X_i$ . Equation (11.10) can be written as

$$LL(\beta_0, \beta_1)=\sum_{i=1}^n Y_i(\beta_0+\beta_1X_i)-\sum_{i=1}^n \ln[1+\exp(\beta_0+\beta_1X_i)] \quad (11.11)$$

Taking partial derivatives with respect to  $\beta_0$  and  $\beta_1$  and equating them to zero, we get the following first-order conditions (Hosmer and Lemeshow, 2000; Kleinbaum and Klein, 2011):

$$\frac{\partial \ln(L(\beta_0, \beta_1))}{\partial \beta_0}=\sum_{i=1}^n Y_i-\sum_{i=1}^n \frac{\exp(\beta_0+\beta_1X_i)}{1+\exp(\beta_0+\beta_1X_i)}=0 \quad (11.12)$$

$$\frac{\partial \ln(L(\beta_0, \beta_1))}{\partial \beta_1}=\sum_{i=1}^n X_i Y_i-\sum_{i=1}^n \frac{X_i \exp(\beta_0+\beta_1X_i)}{1+\exp(\beta_0+\beta_1X_i)}=0 \quad (11.13)$$

Solving Eqs. (11.12) and (11.13) will yield the estimated values of  $\beta_0$  and  $\beta_1$ . However, Eqs. (11.12) and (11.13) do not have closed form solutions; numerical algorithms are used for estimating the parameters. This procedure can be extended when there is more than one explanatory variable in the model.

**EXAMPLE 11.1****Space Shuttle Challenger Data**

Space shuttle orbiter Challenger (Mission STS-51-L) was the 25<sup>th</sup> shuttle launched by NASA on January 28, 1986 (Smith, 1986; Feynman 1988). The Challenger crashed 73 seconds into its flight due to the erosion of O-rings which were part of the solid rocket boosters of the shuttle. Before the launch, the engineers at NASA were concerned about the outside temperature which was very low (the actual launch occurred at 36°F). Data in Table 11.1 shows the O-ring erosion and the launch temperature of the previous shuttle launches, where 'damage to O-ring = 1' implies there was a damage to O-ring and 'damage to O-ring = 0' implies there was no damage to O-ring during that launch. In this case, the outcome is binary – either there is a damage to O-ring or there is no damage to O-ring. We can develop a logistic regression model to predict the probability of erosion of O-ring based on the launch temperature.

**TABLE 11.1** Challenger O-ring erosion data

Flight Number	Launch Temperature	Damage to O-ring	Flight Number	Launch Temperature	Damage to O-ring
STS 1	66	0	STS 41G	78	0
STS 2	70	1	STS 51A	67	0
STS 3	69	0	STS 51C	53	1
STS 4	80	0	STS 51D	67	0
STS 5	68	0	STS 51B	75	0
STS 6	67	0	STS 51G	70	0
STS 7	72	0	STS 51F	81	0
STS 8	73	0	STS 51I	76	0
STS 9	70	0	STS 51J	79	0
STS 41B	57	1	STS 61A	75	1
STS 41C	63	1	STS 61B	76	0
STS 41D	70	1	STS 61C	58	1

The logit function for the example in Table 11.1 is given by

$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = Z = \beta_0 + \beta_1 X_i \quad (11.14)$$

$X_i$  is the launch temperature of  $i^{\text{th}}$  launch. Binary logistic regression output from SPSS is shown in Table 11.2. The values of  $\beta_0$  and  $\beta_1$  are 15.297 and -0.236, respectively.

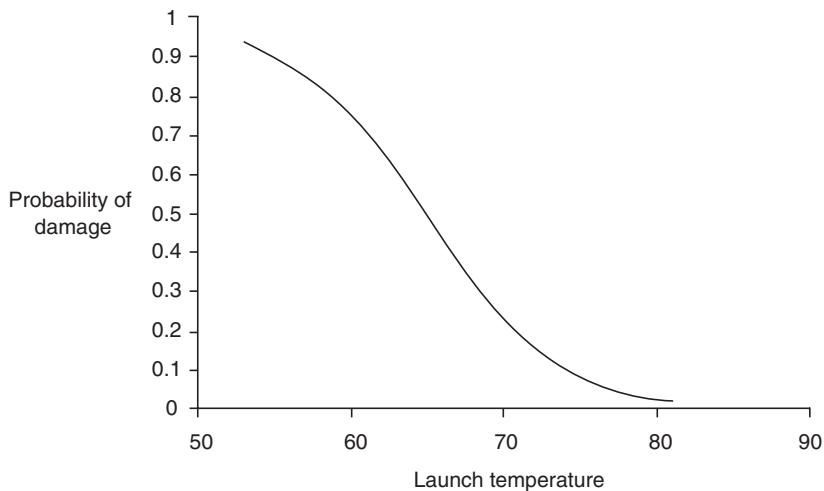
**TABLE 11.2** Logistic regression coefficient for the challenger data in Table 11.1

	<i>B</i> (beta values)	S.E. (Standard error of estimate)	Wald Statistic	<i>df</i>	Sig. ( <i>p</i> -value)
Launch Temperature	-0.236	0.107	4.832	1	0.028
Constant	15.297	7.329	4.357	1	0.037

Using Eq. (11.1), the probability of damage to O-ring as a function of launch temperature is given by

$$P(Y=1) = \pi(Z) = \frac{\exp(15.297 - 0.236 \times X_i)}{1 + \exp(15.297 - 0.236 \times X_i)} \quad (11.15)$$

The plot of probability of damage to O-ring and launch temperature is shown Figure 11.2 in which the probability of damage decreases as the launch temperature increases. However, we have to test the validity of the model using diagnostic tests before the model can be accepted.

**FIGURE 11.2** Probability of damage to O-ring versus launch temperature.

## 11.4 | INTERPRETATION OF LOGISTIC REGRESSION PARAMETERS

Interpretation of logistic regression parameters is not as simple as in the case of linear regression. Consider the logit function defined below:

$$\ln\left(\frac{P(Y_i=1)}{1-P(Y_i=1)}\right) = \beta_0 + \beta_1 X_i \quad (11.16)$$

Note that the ratio  $\frac{P(Y=1)}{1-P(Y=1)}$  is basically odds. If  $P(Y=1) = 0.80$ , then the odds is

$$\frac{P(Y=1)}{1-P(Y=1)} = \frac{0.8}{0.2} = \frac{4}{1}$$

and is usually expressed as 4:1.

When  $X_1 = 0$ , we get

$$\ln\left(\frac{P(Y_i=1|X_i=0)}{1-P(Y_i=1|X_i=0)}\right) = \beta_0 \quad (11.17)$$

When  $X_1 = 1$ , we get

$$\ln\left(\frac{P(Y_i=1|X_i=1)}{1-P(Y_i=1|X_i=1)}\right) = \beta_0 + \beta_1 \quad (11.18)$$

The value of  $\beta_1$  from Eqs. (11.17) and (11.18) is given by

$$\beta_1 = \ln\left(\frac{\left(\frac{P(Y_i=1|X_i=1)}{1-P(Y_i=1|X_i=1)}\right)}{\left(\frac{P(Y_i=1|X_i=0)}{1-P(Y_i=1|X_i=0)}\right)}\right) \quad (11.19)$$

That is,  $\beta_1$  is the change in log-natural of ratio of odds. Alternatively,  $\exp(\beta_1)$  or  $e^{\beta_1}$  is the change in the odds ratio [Eq. (11.20)]:

$$e^{\beta_1} = \left( \frac{\left(\frac{P(Y=1|X_1=1)}{1-P(Y=1|X_1=1)}\right)}{\left(\frac{P(Y=1|X_1=0)}{1-P(Y=1|X_1=0)}\right)} \right) \quad (11.20)$$

An easier interpretation of the regression coefficient,  $\beta_1$ , in logistic regression is that

1. When  $\beta_1$  is positive,  $P(Y=1)$  increases as the value of the predictor  $X_i$  increases.
2. When  $\beta_1$  is negative,  $P(Y=1)$  decreases as the value of the predictor  $X_i$  increases.

In Table 11.2, the coefficient for launch temperature is negative. Thus  $P(Y=1)$  will decrease as the launch temperature increases.

## 11.5 | LOGISTIC REGRESSION MODEL DIAGNOSTICS

We have to carry out diagnostics tests before a binary logistic regression model can be accepted for deployment. Statistical significance of logistic regression model is checked using likelihood ratio test (Omnibus test), Wald's test and Hosmer and Lemeshow test for deployment. We discuss some of these tests below:

1. **Omnibus test:** Omnibus tests are generic statistical tests used for checking whether the variance explained by the model is more than the unexplained variance. For example, in MLR model,  $F$ -test is an omnibus test.  $F$ -test in MLR compares the explained variation with unexplained variation. In the case of logistic regression, this is achieved using likelihood ratio test. Likelihood ratio test usually compares two likelihood functions: one without any independent variable and the other with independent variables. Likelihood ratio test is a chi-square test with degrees of freedom equal to the number of independent variables in the model. Note that a likelihood test can also be used for model comparison; in that case, the degrees of freedom will be the difference in the number of independent variables between the models.
2. **Wald's test:** Wald's test is used for checking whether an individual explanatory variable is statistically significant. Wald's test is a chi-square test.
3. **Hosmer–Lemeshow test:** It is a chi-square goodness of fit test for binary logistic regression.
4. **Pseudo  $R^2$ :** Pseudo  $R^2$  is a measure of goodness of the model. It is called pseudo  $R^2$  because it does not have the same interpretation of  $R^2$  as in the MLR model.

### 11.5.1 | Omnibus Test (Likelihood Ratio Test)

Omnibus tests are generic class of tests in statistics which check whether the explained variance in the model is significantly higher than the unexplained variation. In logistic regression model, likelihood ratio test is used as the test for checking the statistical significance of the overall model. The null and alternative hypothesis for a logistic regression model with  $k$  independent variables are given by

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_A: \text{Not all } \beta\text{s are zero.}$$

We will be using Example 11.1 to explain steps involved in the likelihood ratio test (Example 10.1 has only one independent variable). The log likelihood function for binary logistic regression model is given by

$$LL = \sum_{i=1}^n Y_i \ln[\pi(Z)] + \sum_{i=1}^n (1 - Y_i) [\ln(1 - \pi(Z))] \quad (11.21)$$

Let  $N$  be the total number of observations in the data and let  $N_0$  be the number of 0s (or negatives) and  $N_1$  be the number of 1s (or positives) in the data set. Then, in the absence of any logistic regression model (i.e., model without any independent variable), the estimates of  $P(Y = 1)$  and  $P(Y = 0)$  are given by

$$P(Y = 1) = \pi(Z) = \frac{N_1}{N}$$

$$P(Y = 0) = 1 - \pi(Z) = \frac{N_0}{N}$$

Substituting the values of  $\pi(Z)$  in the log-likelihood function in Eq. (11.21) and multiplying with  $-2$  (to make it chi-square distribution), we get

$$-2LL_0 = -2[N_0 \ln(N_0 / N) + N_1 \ln(N_1 / N)] \quad (11.22)$$

In Eq. (11.22),  $-2LL_0$  implies  $-2$  log-likelihood function with no independent variables. The probability values  $\pi(Z)$  and  $1 - \pi(Z)$  are estimated using frequency of positives and negatives in the data set.

In Table 11.1, there are 24 observations and out of which 7 are positives and the remaining 17 are negatives. The  $-2LL_0$  (also known as null deviance) for the challenger data is given by

$$-2LL_0 = -2[N_0 \ln(N_0 / N) + N_1 \ln(N_1 / N)] = -2[17 \times \ln(17 / 24) + 7 \times \ln(7 / 24)] = 28.975$$

We can calculate the value of  $-2LL_M$  using Eq. (11.15) and Eq. (11.21), which gives the  $-2LL$  for the model (after including the predictor variables). The  $-2LL_M$  for the challenger crash data is shown in Table 11.3.  $-2LL_M$  is calculated iteratively since the logistic regression parameters are calculated iteratively.

**TABLE 11.3**  $-2LL_M$  function value after adding temperature

Iteration	$-2$ Log Likelihood	Coefficients	
		Constant	Launch Temperature
6	20.371	15.297	-0.236

The difference between  $-2LL_0$  and  $-2LL_M$  is  $(28.975 - 20.371) = 8.603$ , which indicates the change in the likelihood function.  $-2LL_M$  is called residual deviance.

$$-2LL_0 - (-2LL_M) = -2 \ln\left(\frac{L_0}{L_M}\right) = 8.603$$

Since  $-2 \ln(L_0 / L_M)$  is a ratio of two likelihood functions, it is called the likelihood ratio. Wilks (1938) proved that  $-2 \ln(L_0 / L_M)$  follows an approximate chi-square distribution. The statistic  $-2 \ln(L_0 / L_M)$  calculates reduction in deviance which is similar to reduction in SSE after adding a variable in case of MLR model. Larger value of deviance implies a poorer fit. We want to reduce the null deviance as much as possible to make the predicted and observed values as close as possible. Deviance also helps user for model selection (similar to partial F-test MLR); a model with lower residual deviance is preferred. The likelihood ratio test output from SPSS for Example 11.1 is shown in Table 11.4.

**TABLE 11.4** Omnibus tests of model coefficients

	Chi-square	df	Sig. (p-value)
Model	8.603	1	0.003

From Table 11.4, it is evident that the likelihood ratio test is statistically significant (since the  $p$ -value is 0.003). That is, the predictor variable, launch temperature, is statistically significant.

### 11.5.2 | Wald's test

Wald's test is used for checking statistical significance of individual predictor variables (equivalent to  $t$ -test in MLR model). The null and alternative hypotheses for Wald's test are:

$$\begin{aligned} H_0: \beta_i &= 0 \\ H_1: \beta_i &\neq 0 \end{aligned}$$

Wald's test statistic is given by

$$W = \left[ \frac{\hat{\beta}_i}{S_e(\hat{\beta}_i)} \right]^2 \quad (11.23)$$

Wald's test is a chi-square test with degrees of freedom = 1. In Table 11.2, the Wald's test statistic value is 4.832 and the corresponding  $p$ -value is 0.028. Since the  $p$ -value is less than 0.05, we reject the null hypothesis. That is, the variable 'launch temperature' has a statistically significant relationship on O-ring failure. The confidence interval for the logistic regression coefficient at  $(1 - \alpha)100\%$  is given by

$$\hat{\beta} \mp |Z_{\alpha/2}| \times S_e(\hat{\beta})$$

where  $S_e(\hat{\beta})$  is the standard error of estimate of  $\hat{\beta}$ .

### 11.5.3 | Hosmer–Lemeshow Test

Hosmer–Lemeshow (H–L) is a chi-square goodness of fit test used for checking the goodness of logistic regression model (Hosmer and Lemeshow, 2000). The H–L test is constructed by dividing the data set into 10 groups (deciles). The H–L test checks whether the observed and expected frequencies in each group are equal. The null and alternative hypotheses in H–L test are

$$\begin{aligned} H_0: \text{The logistic regression model fits the data} \\ H_1: \text{The logistic regression model does not fit the data} \end{aligned}$$

The H–L test statistic is given by

$$H = \sum_{k=1}^G \left[ \frac{(O_k - E_k)^2}{N_k \pi_k (1 - \pi_k)} \right] \quad (11.24)$$

In Eq. (11.24),  $O_k$  is the observed frequency in group  $k$ ,  $E_k$  is the expected frequency in group  $k$ ,  $N_k$  is the number of observations in group  $k$  and  $\pi_k$  is the group mean. The Hosmer–Lemeshow output for the challenger data is given in Table 11.5.

**TABLE 11.5** Hosmer and Lemeshow test

Step	Chi-square	df	Sig.
1	9.922	8	0.271

Since the  $p$ -value in Table 11.5 is 0.271, we retain the null hypothesis, that is the logistic regression model fits the data.

#### 11.5.4 | Pseudo $R^2$

It is not possible to calculate  $R^2$  as in the case of continuous dependent variable in a logistic regression model. However, many pseudo  $R^2$  values are used which compare the intercept-only model to the model with independent variables. In this section, we will discuss Cox and Snell  $R^2$  and Nagelkerke's  $R^2$  which are popular pseudo  $R^2$  measures.

#### Cox and Snell $R^2$

Cox and Snell  $R^2$  is given by (Cox and Snell, 1989)

$$R^2 = 1 - \left\{ \frac{L(\text{Intercept-only model})}{L(\text{Full model})} \right\}^{2/N} \quad (11.25)$$

where  $N$  is the sample size.  $L(\text{Full model})$  is the likelihood function which is used for calculating the class probability. The maximum value of Cox and Snell  $R^2$  may not become 1. The maximum value of Cox and Snell  $R^2$  is  $1 - \{L(\text{Intercept-only model})\}^{2/N}$ .

#### Nagelkerke's $R^2$

Nagelkerke  $R^2$  is an adjustment over Cox and Snell  $R^2$ , so that the maximum value Pseudo  $R^2$  is 1. Nagelkerke  $R^2$  is given as follows (Nagelkerke, 1991):

$$R^2 = \frac{1 - \left\{ \frac{L(\text{Intercept-only model})}{L(\text{Full model})} \right\}^{2/N}}{1 - \{L(\text{Intercept-only model})\}^{2/N}} \quad (11.26)$$

### 11.6 | CLASSIFICATION TABLE, SENSITIVITY, AND SPECIFICITY

The primary objective of logistic regression is to solve classification problems based on the predicted probability values. The output from a logistic regression model is the class probability  $P(Y = 1)$ . Based on the value of  $P(Y = 1)$ , the decision maker has to classify the observation as belonging to either class 1 (positive) or class 0 (negative). To classify the observations, the decision maker has to first decide the classification cut-off probability  $P_c$ . Whenever the predicted probability of an observation,  $P(Y_i = 1)$ , is less than the classification cut-off probability,  $P_c$ , then the observation is classified as negative ( $Y_i = 0$ ) and if the predicted probability is greater than or equal to  $P_c$ , then the observation is classified as positive ( $Y_i = 1$ ). That is

$$Y_i = \begin{cases} 0 & \text{if } P(Y_i = 1) < P_c \\ 1 & \text{if } P(Y_i = 1) \geq P_c \end{cases}$$

Note that the value of classification cut-off probability is a decision to be taken by the decision maker to enable classification of observations in the data set. Later in the chapter we will be discussing the different approaches for deriving optimal classification cut-off probability. Many software tools use a default classification cut-off probability of 0.5. The classification table in a logistic regression model output is a table that provides accuracy of the logistic regression model (accuracy of classifying positives and negatives) for a chosen classification cut-off probability. Table 11.6 shows the classification table (also known as error matrix or confusion matrix) for the Example 11.1 for a cut-off probability of 0.5.

**TABLE 11.6** Classification Table<sup>a</sup>

Observed		Predicted		Percentage Correct
		Damage to O-ring		
Step 1	Damage to O-ring	0 (Negative)	17 (TN)	0 (FP)
		1 (Positive)	3 (FN)	4 (TP)
	Overall Percentage			87.5

<sup>a</sup>The cut value is 0.500.

The Challenger data had 17 no-damage (negative) cases and 7 damage (positive) cases. The probability of damage to O-ring is calculated using the logistic function in Eq. (11.15). When the probability is less than 0.5, the observation is classified as negative (coded as 0) and when the probability is greater than or equal to 0.5, the observation is classified as positive (coded as 1). For the classification cut-off probability value of 0.5, the model has classified all 17 negatives (coded as 0) as negatives and 4 positives (coded 1) as positives and remaining 3 positives as negatives. The accuracy of classifying negatives is 100%, whereas the accuracy of classifying positives is 57.1% when the classification cut-off probability is 0.5. The overall accuracy of the logistic regression model is 87.5% (3 observations are misclassified out of 24 observations). The classification accuracy would depend on the cut-off probability. Classification table for the cut-off probability of 0.2 for example 11.1 is shown in Table 11.7.

**TABLE 11.7** Classification Table<sup>a</sup>

Observed		Predicted		Percentage Correct
		Damage to O-ring		
Step 1	Damage to O-ring	0	9	52.9
		1	1	85.7
	Overall Percentage			62.5

<sup>a</sup>The cut value is 0.200.

When the cut-off probability is 0.2, the overall accuracy decreases to 62.5; however, the accuracy of predicting positives increases from 57.1 to 85.7. However, the accuracy of predicting negatives has decreased to 52.9% from 100% and the overall accuracy has decreased to 62.5% from 87.5%.

### 11.6.1 | Accuracy Paradox

The accuracy paradox in classification problem states that a model with higher overall accuracy may not be a better model. For example, in the case of Challenger data, when the classification cut-off probability is 0.5, the overall accuracy is 87.5 and when the classification cut-off probability is 0.2, the overall accuracy is only 62.5. However, given the context, we need higher accuracy in predicting positive classes ( $Y_i = 1$ ) than negative classes ( $Y_i = 0$ ). The accuracy in predicting positive classes when classification cut-off is 0.5 is 57.1, whereas the corresponding accuracy for classification cut-off probability of 0.2 is 85.7. In this case, classification cut-off 0.2 is better than 0.5 although the overall classification accuracy is lesser for classification cut-off probability 0.2. In classification problems, the model selection cannot be based on overall accuracy.

### 11.6.2 | Sensitivity, Specificity, and Precision

In logistic regression, the model performance is often measured using concepts such as sensitivity, specificity and precision. The ability of the model to correctly classify positives and negatives are called sensitivity and specificity, respectively. The terminologies sensitivity and specificity originated in medical diagnostics. In medical diagnostics, **sensitivity** (also known as true positive rate) measures the ability of a diagnostic test to identify disease if it is present in a patient (test positive). That is

$$\text{Sensitivity} = P(\text{diagnostic test is positive} \mid \text{patient has disease})$$

In generic case

$$\text{Sensitivity} = P(\text{model classifies } Y_i \text{ as positive} \mid Y_i \text{ is positive})$$

Sensitivity is calculated using the following equation:

$$\text{Sensitivity} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Negative (FN)}} \quad (11.27)$$

where True Positive (TP) is the number of positives correctly classified as positives by the model and False Negative (TN) is positives misclassified as negative by the model. Sensitivity is also called as **recall**.

**Specificity** is the ability of the diagnostic test to correctly classify the test as negative when the disease is not present. That is:

$$\text{Specificity} = P(\text{diagnostic test is negative} \mid \text{patient has no disease})$$

In general:

$$\text{Specificity} = P(\text{model classifies } Y_i \text{ as negative} \mid Y_i \text{ is negative})$$

Specificity can be calculated using the following equation:

$$\text{Specificity} = \frac{\text{True Negative (TN)}}{\text{True Negative (TN)} + \text{False Positive (FP)}} \quad (11.28)$$

where True Negative (TN) is number of the negatives correctly classified as negatives by the model and False Positive (FP) is number of negatives misclassified as positives by the model.

In the Challenger example (Example 11.1), when the cut-off probability is 0.5, the sensitivity and specificity are given by (Table 11.6)

$$\text{Sensitivity} = \frac{4}{4+3} = \frac{4}{7} = 0.571 \text{ or } 57.1\%$$

$$\text{Specificity} = \frac{17}{17+0} = \frac{17}{17} = 1 \text{ or } 100\%$$

When the cut-off probability is 0.2, the sensitivity and specificity are 82.5% and 52.9%, respectively (Table 11.7). When the cut-off is decreased to 0.5 from 0.2, the sensitivity increases from 57.1% to 82.5%; however, the specificity has decreases from 100% to 52.9%. That is, there is tradeoff between sensitivity and specificity in most classification problems. The decision maker has to consider the tradeoff between sensitivity and specificity to arrive at an optimal cut-off probability.

Precision measures the accuracy of positives classified by the model.

$$\text{Precision} = P(\text{patient has disease} \mid \text{diagnostic test is positive})$$

$$\text{Precision} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Positive (FP)}} \quad (11.29)$$

**F-Score (F-Measure)** is another measure used in binary logistic regression that combines both precision and recall (harmonic mean of precision and recall) and is given by

$$F - \text{Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (11.30)$$

### 11.6.3 | Concordant and Discordant Pairs

The concepts of concordant and discordant pairs are measures used to assess the logistic regression model performance. Using Eq. (11.15), we predict the probability of damage to O-ring for the data in Table 11.1. The predicted probability of damage to O-ring using Eq. (11.15) is provided in Table 11.8.

**TABLE 11.8** Challenger crash data – predicted probability using logistic regression model

S. No.	Flight Number	Launch Temperature	Damage to O-ring	Predicted Probability
1	STS 1	66.00	0	0.43
2	STS 2	70.00	1	0.23
3	STS 3	69.00	0	0.27
4	STS 4	80.00	0	0.03
5	STS 5	68.00	0	0.32
6	STS 6	67.00	0	0.37
7	STS 7	72.00	0	0.15
8	STS 8	73.00	0	0.13

**TABLE 11.8** Challenger crash data – predicted probability using logistic regression model—Continued

S. No.	Flight Number	Launch Temperature	Damage to O-ring	Predicted Probability
9	STS 9	70.00	0	0.23
10	STS 41B	57.00	1	0.86
11	STS 41C	63.00	1	0.61
12	STS 41D	70.00	1	0.23
13	STS 41G	78.00	0	0.04
14	STS 51A	67.00	0	0.37
15	STS 51C	53.00	1	0.94
16	STS 51D	67.00	0	0.37
17	STS 51B	75.00	0	0.08
18	STS 51G	70.00	0	0.23
19	STS 51F	81.00	0	0.02
20	STS 51I	76.00	0	0.07
21	STS 51J	79.00	0	0.03
22	STS 61A	75.00	1	0.08
23	STS 61B	76.00	0	0.07
24	STS 61C	58.00	1	0.83

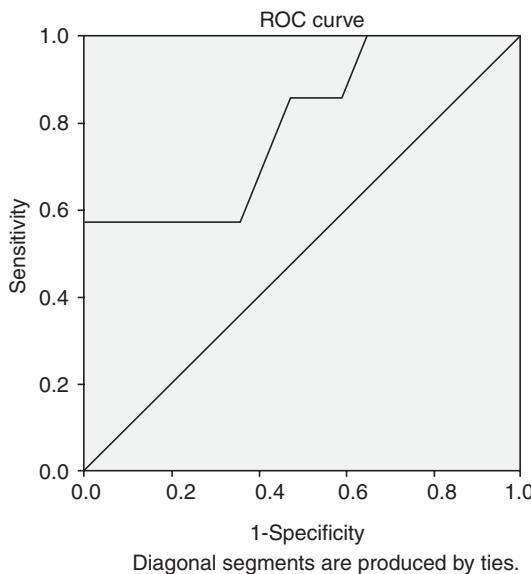
Consider observations 1 and 2 in Table 11.8. The probabilities of damage to O-ring for observations 1 and 2 (STS 1 and STS 2) are 0.43 and 0.23, respectively. The launch temperature for STS 1 is 66°F and STS 2 is 70°F. If we are given these two observations, then there is no cut-off probability that can classify them correctly. For example, if we choose a cut-off probability which is less than 0.23, then STS 1 will be misclassified since predicted probability of STS 1 is 0.43. If we choose a classification cut-off probability between 0.23 and 0.43, then both of them will be misclassified, since probability of STS 2 is 0.23 (whereas the corresponding  $Y = 1$ ) and probability of STS 1 is 0.43 (the value of  $Y = 0$ ). If the cut-off probability is greater than 0.43, the STS 2 will be misclassified. Thus, if we select STS 1 (for which  $Y = 0$ ) and STS 2 (for which  $Y = 1$ ), there is no cut-off probability that can classify both positive ( $Y = 1$ ) and negative ( $Y = 0$ ) correctly. Such pairs are called **Discordant Pairs**. That is, a pair of positive and negative observations for which the model has no cut-off probability to classify both of them correctly are called discordant pairs.

Consider observations STS 9 (launch temperature is 70°F and  $Y = 0$ ) and STS 41B (launch temperature is 57°F and  $Y = 1$ ). Predicted probability of damage to O-ring for STS 9 is 0.23 and STS 41B is 0.86. If we use a classification cut-off probability between 0.23 and 0.86, then we will classify STS 9 ( $Y = 0$ ) and STS 41B ( $Y = 1$ ) correctly. Such pairs are called **Concordant Pairs**. That is, a pair of positive and negative observations for which the model has a cut-off probability to classify both of them correctly are called concordant pairs. A logistic regression model with high proportion of concordant pairs is preferred.

In general, let  $(a_i, b_i)$  be dyadic data set such that  $a_i \in P$  (where  $P$  is the set of positive observations) and  $b_i \in N$  (where  $N$  is the set of negative observations). Then, the concordant pair is a pair such that the logistic regression model can correctly classify using a cut-off probability, whereas discordant pair is a pair for which the model does not have a cut-off probability to correctly classify.

### 11.6.4 | Receiver Operating Characteristics (ROC) Curve

Receiver operating characteristic (ROC) curve can be used to understand the overall worth of a logistic regression model (and, in general, for classification models). The term has its origin in electrical engineering when electrical signals were used for predicting enemy objects (such as submarines and aircraft) during World War II. ROC curve is a plot between sensitivity (true positive rate) in the vertical axis and  $1 - \text{specificity}$  (false positive rate) in the horizontal axis. In Section 11.6.2, we saw that when the classification cut-off probability is changed, the sensitivity and specificity are likely to change. The ROC curve for the challenger crash data is shown in Figure 11.3.



**FIGURE 11.3** ROC curve for challenger crash data.

In Figure 11.3, the diagonal line represents the case of not using a model (no discrimination between positive and negative); the area below the diagonal line is equal to 0.5 (it is a right-angle triangle, area of right-angle triangle is  $(1/2)ab$ , where  $a$  and  $b$  are the length of the sides; in this case  $a = b = 1$ ). When we use a model, the sensitivity and/or specificity is likely to change. The line above the diagonal line captures how sensitivity and  $1 - \text{specificity}$  change when the cut-off probability is changed. The area under the ROC curve (AUC) for the challenger crash data is given in Table 11.9. The AUC is the proportion of the concordant pairs in the data. Model with higher AUC is preferred and AUC is frequently used for model selection.

**TABLE 11.9** Area under the curve

Area	Std. Error <sup>a</sup>	Asymptotic Sig. <sup>b</sup>	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.794	0.107	0.026	0.585	1.000

The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased. <sup>a</sup>Under the nonparametric assumption <sup>b</sup>Null hypothesis: true area = 0.5.

**IMPORTANT**

AUC is the proportion of concordant pairs in the data if the model is used for classification. AUC is one of the criteria used for final model selection; higher AUC is assumed to be a better model.

For challenger crash data, the AUC is 0.794. The area under the ROC curve can be interpreted as follows:

1. If we use the logistic regression model, then there will be 79.4% concordant pairs and 20.6% discordant pairs.
2. For a randomly selected pair of positive and negative observations, probability of correctly classifying them is 0.794.
3. For a randomly selected positive and negative observations, the following relationship is valid:

$$P[P(\text{Positive}) > P(\text{Negative})] = 0.794 \quad (11.31)$$

That is, for a randomly selected one positive observation and one negative observation, say  $Y_k = 1$  and  $Y_j = 0$ , AUC of 0.794 implies that  $P(P(Y_k = 1) > P(Y_j = 1)) = 0.794$ .

As a thumb rule, AUC of at least 0.7 is required for practical application of the model. AUC of greater than 0.9 implies an outstanding model. Caution should be used while selecting models based on AUC, especially when the data is imbalanced (that is, data set which has less than 10% positives). In case of imbalanced data sets, the AUC may be very high (greater than 0.9); however, either sensitivity or specificity values may be poor.

### 11.6.5 | Area Under ROC Curve (AUC), Lorenz Curve, and Gini Coefficient

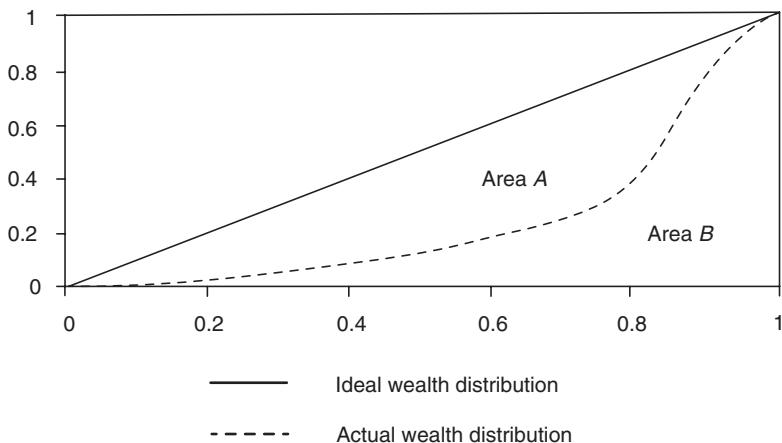
Max O Lorenz, an American economist, developed a wealth distribution plot (Figure 11.4) to quantify the discrimination among citizens of a country based on the total wealth in a society (Gastwirth, 1971). Horizontal axis in Figure 11.4 denotes the proportion of population and the vertical axis denotes proportion of wealth. The diagonal line in Figure 11.4 indicates that there is no wealth discrimination, that is, the wealth is equally distributed among the population. Curve A indicates the observed wealth distribution as a function of population proportion. As the size of area A in the Lorenz curve increases, the discrimination increases; similarly as the size of B increases, the wealth discrimination decreases. The discrimination is usually measured using Gini coefficient and is given by

$$\text{Gini coefficient} = \left( \frac{\text{Area } A}{\text{Area } A + \text{Area } B} \right) \quad (11.32)$$

Note that, in this case,  $\text{Area } A + \text{Area } B = 1/2$ , since it is a right-angle triangle in which the length of the sides is equal to 1. Substituting the value of  $\text{Area } A + \text{Area } B (= 1/2)$  in Eq. (11.32), we get the Gini coefficient as  $2A$ . We can connect the area under ROC curve (AUC) with Gini coefficient. In ROC curve, the area A is above the diagonal line. Thus,  $\text{AUC} = (1/2) + \text{Area } A$  or  $2\text{AUC} = 1 + 2 \text{Area } A = 1 + \text{Gini Coefficient}$ . That is,

$$\text{Gini coefficient} = 2\text{AUC} - 1 \quad (11.33)$$

Gini coefficient is useful in logistic regression model building for variable selection since a variable with high Gini coefficient indicates that the variable is able to discriminate the values of response variable well.



**FIGURE 11.4** Lorenz curve.

When there are a large number of independent variables, Gini coefficient may be used as a strategy to shortlist (or rank) the variables for LR model building. In such cases we develop univariate LR models (models with single independent variable) and rank the independent variables based on their Gini coefficient (or AUC).

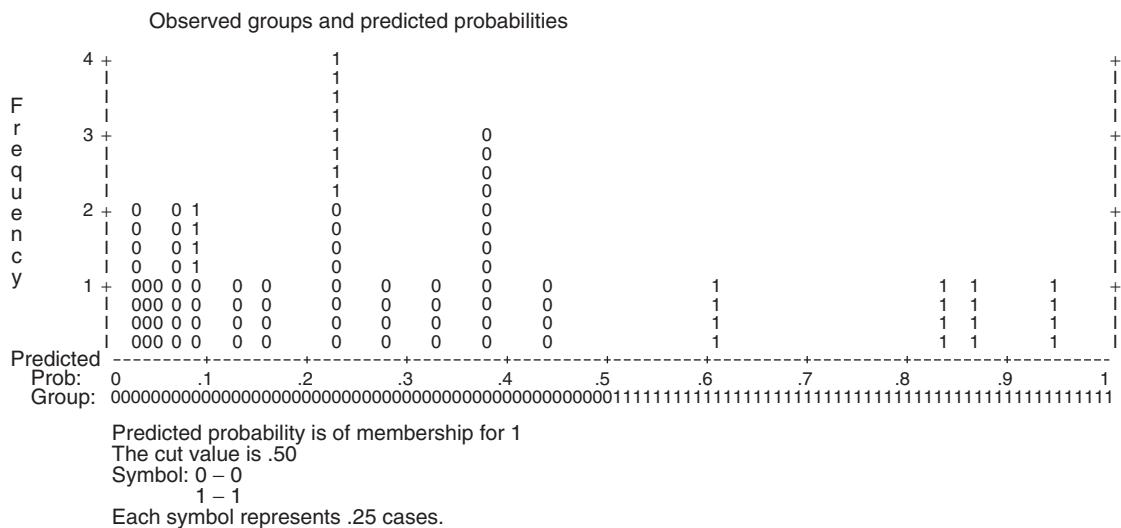
## 11.7 | OPTIMAL CUT-OFF PROBABILITY

While using logistic regression model, one of the decisions that a data scientist has to make is to choose the right classification cut-off probability ( $P_c$ ). The overall accuracy, sensitivity and specificity will depend on the chosen cut-off probability. The following three methods are used for selecting the cut-off probability.

1. Classification plot
2. Youden's Index
3. Cost based approach

### 11.7.1 | Classification Plot for Selection of Cut-off Probability

Classification plot is a plot between the predicted probability on horizontal axis and the corresponding frequencies of the observations using LR model. Classification plot for challenger crash data obtained using SPSS is shown in Figure 11.5.



**FIGURE 11.5** Classification plot for challenger crash data for logistic regression.

In Figure 11.5, we can see that there is no negative (coded as  $Y = 0$ ) beyond the probability of 0.5, whereas, there are positives even at probability value of approximately 0.08 (refer to Table 11.8). So, if we have to ensure that all the positives are correctly identified, then we have to set a classification cut-off probability which is less than 0.08 (say 0.05). In Figure 11.5, each symbol represents approximately 0.25 cases. The classification table when we use a cut-off probability of 0.05 is shown in Table 11.10.

**TABLE 11.10** Classification Table<sup>a</sup>

		Predicted			
		Observed		Predicted	
				Damage to O-ring	
				0 (Negative)	1 (Positive)
Step 1	Damage to O-ring	0 (Negative)		4 (TN)	13 (FP)
		1 (Positive)		0 (FN)	7 (TP)
Overall Percentage					45.8

<sup>a</sup>The classification cut-off value is 0.050.

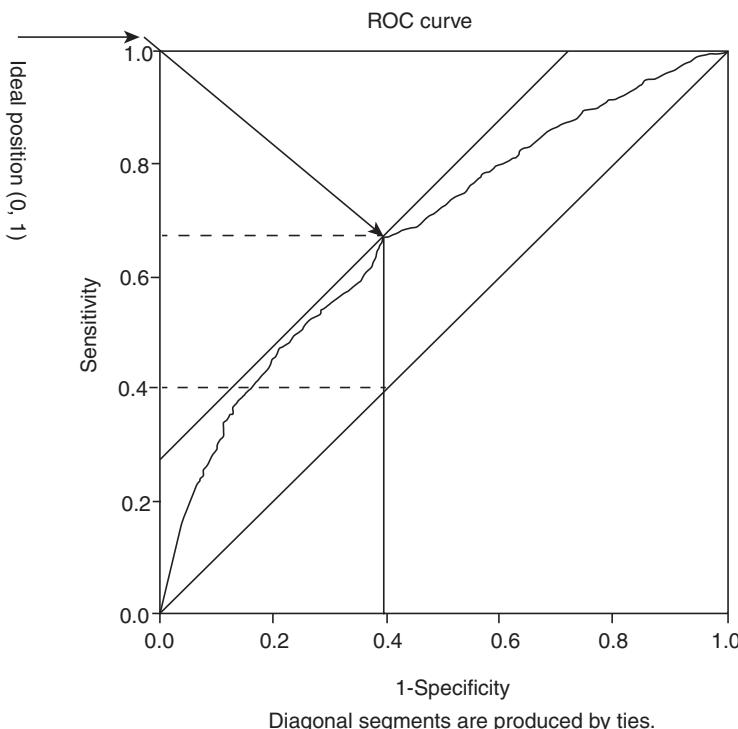
At a cut-off value of 0.05, we are able to classify all positives ( $Y = 1$ ) correctly (sensitivity = 100%), whereas only 4 out of 17 negatives ( $Y = 0$ ) are correctly classified, that is the specificity is only 23.5%. Although, classification plot may help us to understand how sensitivity and specificity are likely to change when we change the classification cut-off probability, arriving at an optimal cut-off is difficult using classification plot.

### 11.7.2 | Youden's Index for Optimal Cut-Off Probability

Sensitivity and specificity change when we change the cut-off probability. Youden's Index (Youden, 1950) is a classification cut-off probability,  $P_c^*$ , for which the following function is maximized (also known as J statistic):

$$\text{Youden's Index} = \text{J Statistic} = \underset{p}{\text{Max}} [\text{Sensitivity}(p) + \text{Specificity}(p) - 1] \quad (11.34)$$

The logic for using Youden's Index for finding optimal cut-off probability can be explained using the area under the ROC curve. Consider the ROC curve shown in Figure 11.6. In this figure, the coordinate (0, 1) implies sensitivity = specificity = 1 which is the ideal model that we would like to use. The ROC curve provides information regarding how sensitivity and specificity change when the classification cut-off probability changes. The point on the ROC curve which is at minimum distance from coordinate (0, 1) (or which is at the maximum distance from the diagonal line) will give us the best cut-off probability. Youden's Index is the classification cut-off probability value for which the distance from the diagonal line to the ROC curve is maximum. Using simple algebra we can show that the optimal cut-off is given by the one that maximizes right-hand side of Eq. (11.34). Note that the maximum distance is a line parallel to the diagonal line which is also tangent to the ROC curve.



**FIGURE 11.6** Youden's index.

We can calculate Youden's Index by incrementally changing the cut-off probability and calculating the corresponding sensitivity + specificity - 1. Youden's Index for challenger crash data is shown in Table 11.11. The maximum value of Youden's Index is 0.571 which occurs for cut-off probability values of 0.5 and 0.6 (in fact for all the values between 0.5 and 0.6). But given the context, one would expect 100% accuracy for sensitivity in this case. We use Youden's Index only when both sensitivity and specificity are equally important. When sensitivity and specificity are not equally important then we use cost-based approach for cut-off probability (discussed in the next section).

**TABLE 11.11** Youden's index for challenger crash data

Cut-off	Sensitivity	Specificity	Youden's Index
0.05	1	0.235	0.235
0.1	0.857	0.412	0.269
0.2	0.857	0.529	0.386
0.3	0.571	0.706	0.277
0.4	0.571	0.941	0.512
0.5	0.571	1	0.571
0.6	0.571	1	0.571
0.7	0.429	1	0.429
0.8	0.429	1	0.429
0.9	0.143	1	0.143
1	0	1	0

### 11.7.3 | Cost-Based Cut-Off Probability

In cost-based approach, we assign penalty cost for misclassification of positives and negatives. Assume that cost of misclassifying negative (0) as positive (1) is  $C_{01}$  and cost of misclassifying positive (1) as negative (0) is  $C_{10}$  as shown in Table 11.12.

**TABLE 11.12** Cost of misclassification

Observed	Classified	
	0	1
0	—	$C_{01}$
1	$C_{10}$	—

The optimal cut-off probability is the one which minimizes the total penalty cost and is given by

$$\text{Min}_p [C_{01}P_{01} + C_{10}P_{10}] \quad (11.35)$$

Here  $P_{01}$  and  $P_{10}$  are the probability of classifying negative as positive and positive as negative, respectively for a cut-off probability  $p$ .  $C_{01}P_{01} + C_{10}P_{10}$  is the expected total penalty cost of misclassification.

In Eq. (11.35) we try to find the cut-off probability  $p$  that minimizes the expected total penalty cost. Cut-off probability based on penalty cost is the most preferred in cases such as credit rating.

## 11.8 | VARIABLE SELECTION IN LOGISTIC REGRESSION

In this section we discuss two approaches that can be used for developing logistic regression models by selecting variables automatically. Such procedures ensure that only statistically significant variables at a significance value of  $\alpha$  are included in the model. There are several criteria used for variable selection in logistic regression (Hosmer *et al.* 1989). In this section we discuss variable selection based on likelihood ratio test and Wald's test.

### 11.8.1 | Forward LR (Likelihood Ratio)

In Forward LR (Lawless and Singhal, 1987), at each step one variable is added to the model. The following steps are used in building logistic regression model using forward LR selection method.

#### STEP 1

---

Start with no variables in the model. Set  $i = 0$ .

---

#### STEP 2

For each independent variable, calculate the difference between  $-2LL_i$  and  $-2LL_{i+1}$  value. When  $i = 0$ , we will calculate the difference between  $-2LL_0$  (model without a variable) and  $-2LL_1$  (model with one variable). Add the variable with highest difference in  $-2LL_i$  and  $-2LL_{i+1}$  if the  $p$ -value after adding the variable is statistically significant under likelihood ratio test (omnibus test) at a significance level of  $\alpha$ .

---

#### STEP 3

---

Repeat step 2, till all the variables are exhausted or the change in  $-2LL$  is not significant, that is the  $p$ -value after adding a new variable is greater than  $\alpha$ .

---

### 11.8.2 | Forward Selection Wald

In Forward Selection Wald, the variables are entered based on the Wald's test.

#### STEP 1

---

Assume that the data has ' $n$ ' explanatory variables. Develop a univariate LR model and calculate the  $p$ -value of all the variables and add the variable with the smallest  $p$ -value if it is less than  $\alpha$  (significance level).

---

**STEP 2**

Add a new variable with smallest  $p$ -value or largest Wald's statistic value if  $p$ -value (based on the Wald's test) is less than the significance  $\alpha$ .

**STEP 3**

Repeat the procedure till the  $p$ -value becomes greater than  $\alpha$ .

## 11.9 | APPLICATION OF LOGISTIC REGRESSION IN CREDIT RATING

In this section, we will be discussing how logistic regression can be used in credit scoring. For this purpose, we will be using sample data (File name: German Credit Rating.xlsx, the data set has 800 observations and 13 attributes) taken from the 'German Credit Data'<sup>1</sup> available at the University of California, Irvine machine learning repository. Note that the original data set has 20 attributes (predictors) with over 40,000 observations. The response variable  $Y$  takes value 1 (Bad credit) or 0 (Good credit). Predictors used in developing the credit rating model are listed in Table 11.13.

**TABLE 11.13** German credit data attributes

S. No.	Attribute (Data Type)	Description
1.	Credit Classification (Qualitative)	Two classes: Bad (coded as 1), Good (coded as 0)
2.	Checking Account Balance (Categorical)	Four categories: 1. No account, 2. 0 DM (Deutsch Mark), 3. Between 0 and Less than 200 DM and 4. Over 200 DM
3.	Duration (Numerical)	Duration of the credit
4.	Credit Amount (Numerical)	Amount of credit given
5.	Balance in Savings Account (categorical)	Five categories: 1. Unknown, 2. Less than 100 DM, 3. Between 100 and 500 DM, 4. Between 500 and 1000 DM, 5. Over 1000 DM
6.	Employment in Years (Categorical)	Five categories: 1. Unemployed, 2. Less than one year, 3. Between 1 and 4 years, 4. Between 4 and 7 years and 5. More than 7 years
7.	Installment Rate (numerical)	As a percentage of disposable income
8.	Marital Status (categorical)	Four categories: 1. Single, 2. Married, 3. Divorced Male and 4. Divorced Female
9.	Present Resident (Numerical)	Present residence in years
10.	Age (Numerical)	Age of the applicant in years
11.	Other Installments (Binary)	1. Applicant has other installments 2. Applicant has no other installments.
12.	Number of Credits (Numerical)	Number of existing credits in the bank
13.	Job (Categorical)	Four categories: 1. Unskilled, 2. Skilled, 3. Management and 4. Unemployed
14.	Credit History (categorical)	Four categories: 1. all paid duly, 2. Bank paid duly, 3. Delay and 4. Critical

<sup>1</sup> Hans Hofmann, (2015), UCI Machine learning repository, Irvine, CA: University of California, School of Information and Computer Science. (<https://archive.ics.uci.edu/ml/datasets/>)

SPSS output using variable selection technique forward LR (likelihood ratio) is shown in Tables 11.14 and 11.15. In Table 11.14, the classification table is created using a classification cut-off probability of 0.5. The forward LR procedure has included 10 variables that are statistically significant (includes different dummy variables of categorical variables) in the model. The sensitivity and specificity of the logistic regression model after adding 10 variables is 43.6% and 89.8%, respectively. The overall accuracy of the model is 76%. Note that in forward LR, one variable is added at each step and the classification table after steps 1, 2 and 10 are shown in Table 11.14.

**TABLE 11.14** Classification Table<sup>a</sup>

	Observed	Predicted			Percentage Correct
		Credit Rating		0.0	
		1.0			
Step 1	Credit Rating	0.0	549	0	100.0
		1.0	234	0	0.0
Overall Percentage					70.1
Step 2	Credit Rating	0.0	549	0	100.0
		1.0	234	0	0.0
Overall Percentage					70.1
Step 10	Credit Rating	0.0	493	56	89.8
		1.0	132	102	43.6
Overall Percentage					76.0

<sup>a</sup>The cut-off probability value is 0.500.

**TABLE 11.15** Variables in the equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	0 DM	1.044	0.171	37.340	1	0.000	2.841
	Constant	-1.162	0.098	141.872	1	0.000	0.313
Step 2	0 DM	1.800	0.211	72.612	1	0.000	6.048
	Between 0 and 200 DM	1.587	0.209	57.767	1	0.000	4.891
	Constant	-1.918	0.158	147.471	1	0.000	0.147
Step 10	0 DM	1.836	0.247	55.283	1	0.000	6.273
	Between 0 and 200 DM	1.614	0.241	44.775	1	0.000	5.024
	over 200 DM	0.843	0.406	4.317	1	0.038	2.322
	Duration	0.045	0.007	36.017	1	0.000	1.046
	Critical	-0.772	0.233	10.950	1	0.001	0.462
	less than 100	0.429	0.192	5.002	1	0.025	1.536
	Seven years	-0.708	0.253	7.820	1	0.005	0.493
	Install_rate	0.255	0.081	9.848	1	0.002	1.290
	Single	-0.663	0.183	13.086	1	0.000	0.515
	Other installment	0.524	0.194	7.308	1	0.007	1.689
Constant		-3.601	0.403	79.846	1	0.000	0.027

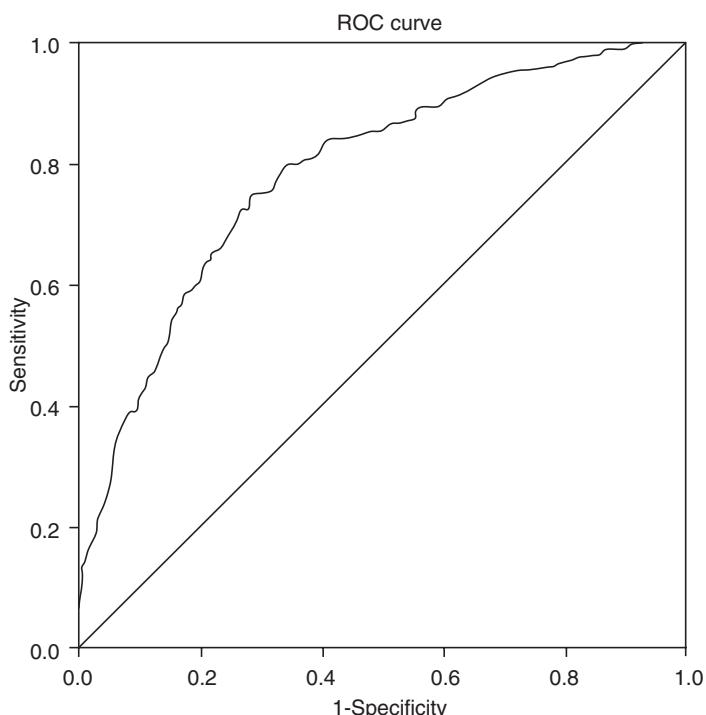
The coefficient values after inclusion of all statistically significant variables are shown in Table 11.15. Note that Table 11.15 includes values of coefficient after steps 1, 2, and 10. In the first step, variable '0 DM (checking account balance = 0 DM)' is added and in the second step variable 'checking account balance between 0 and 200 DM' is added. The probability of bad loan,  $P(Y = 1)$ , is given by

$$P(Y = 1) = \frac{e^Z}{1 + e^Z} \quad (11.36)$$

where

$$\begin{aligned} Z = & -3.601 + 1.836 \times 0 \text{ DM} + 1.614 \times \text{Between 0 and 200 DM} + 0.843 \times \text{Over 200 DM} + 0.045 \times \\ & \text{Duration} - 0.772 \times \text{Critical} + 0.429 \times \text{Less than 100} - 0.708 \times \text{Seven years} + 0.225 \times \text{Install rate} \\ & - 0.663 \times \text{Single} + 0.524 \times \text{Other installment} \end{aligned}$$

Equation (11.36) can be used for calculating the probability of default, and based on the calculated probability, the decision maker may decide whether or not to provide loan to a new applicant. The ROC curve for the model developed is shown in Figure 11.7 and the area under ROC curve is provided in Table 11.16. The area under ROC curve is 0.786, that is, the data set has 78.6% concordant pairs. The classification plot is shown in Figure 11.8.



Diagonal segments are produced by ties.

**FIGURE 11.7** ROC curve for German credit rating sample data.

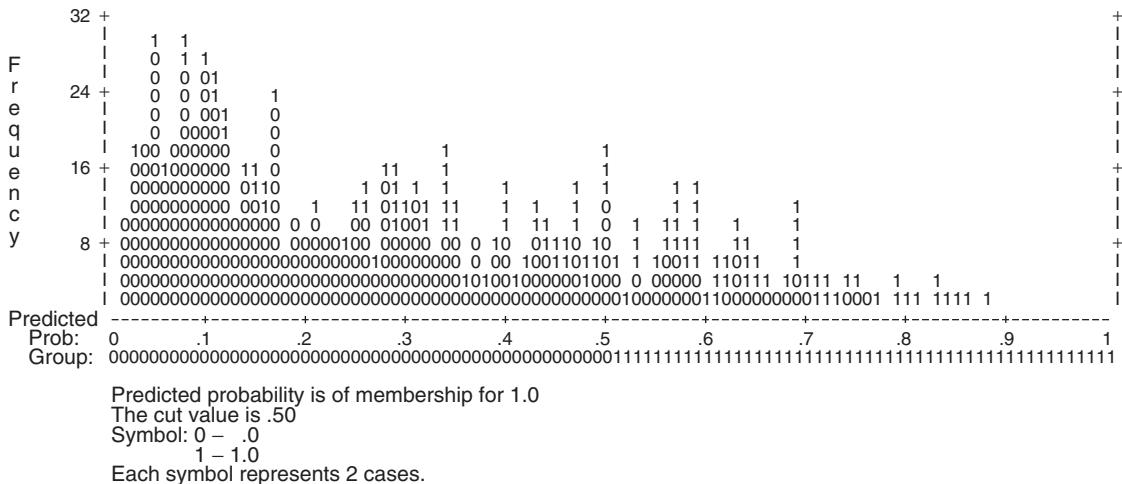
**TABLE 11.16** Area under the curve for german credit data

Area	Std. Error	Asymptotic Sig	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.786	0.017	0.000	0.752	0.820

The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

Step number: 10

### Observed groups and predicted probabilities



**FIGURE 11.8** Classification plot of the logistic regression model for German credit rating data.

## 11.9.1 | CREDIT SCORE USING LOGISTIC REGRESSION

Many credit rating organizations use a range of score (say between two values  $A$  and  $B$ ) called credit score to measure the credit worthiness of a customer. For example, one of the popular credit scores, FICO (Fair, Isaac and Company) has a range between 300 and 850. FICO score is calculated using multiple parameters such as payment history, debt burden, length of credit history, type of credit, etc. A logistic regression model can be used for generating a credit score between any two values  $A$  and  $B$  using

$$\text{Credit Score of Customer} = A + (B - A) \times [1 - P(Y=1)] \quad (11.37)$$

For example, if  $A = 300$  and  $B = 850$ , then the credit score is given by

$$\text{Credit Score of Customer} = 300 + 550 \times [1 - P(Y=1)]$$

When the probability of default  $P(Y = 1) = 1$ , then the credit score is 300 and if the probability of default  $P(Y = 1) = 0$ , then the credit score will be 850.

### 11.9.2 | Youden's Index Calculation

To calculate Youden's index, one has to first calculate the value of sensitivity( $p$ ) + specificity( $p$ ) – 1 for different classification cut-off probability,  $p$ . Table 11.17 provides the values of sensitivity and specificity for different classification cut-off probability values for the model developed for the German credit rating data.

**TABLE 11.17** Youden's index calculation

Classification Cut-off Probability ( $p$ )	Sensitivity( $p$ )	Specificity( $p$ )	[Sensitivity( $p$ ) + Specificity( $p$ ) – 1]
0.05	0.987	0.128	0.115
0.1	0.944	0.313	0.257
0.15	0.885	0.446	0.331
0.2	0.85	0.543	0.393
0.25	0.816	0.612	0.428
<b>0.3</b>	<b>0.752</b>	<b>0.692</b>	<b>0.444</b>
0.35	0.662	0.763	0.425
0.4	0.607	0.805	0.412
0.45	0.538	0.849	0.387
0.5	0.436	0.898	0.334
0.55	0.393	0.922	0.315
0.6	0.278	0.954	0.232
0.65	0.192	0.976	0.168
0.7	0.132	0.993	0.125
0.75	0.073	0.998	0.071
0.8	0.047	1	0.047
0.85	0.013	1	0.013
0.9	0	1	0
0.95	0	1	0

In Table 11.17, the maximum value of sensitivity( $p$ ) + specificity( $p$ ) – 1 is 0.444 and the corresponding classification cut-off probability is 0.30. That is, the value of Youden's Index is 0.444 and the corresponding classification cut-off probability is 0.30, thus the optimal classification cut-off probability using Youden's Index is 0.30.

### 11.9.3 | Classification Cut-Off Based on Penalty Cost

Assume that the penalty cost for misclassification is as defined in Table 11.18. The cost of classifying 1 (positive) as 0 (negative) is 2 times more severe than cost of classifying 0 (negative) as 1 (positive).

**TABLE 11.18** Penalty cost for misclassification

Observed	Predicted	
	0	1
0	0	1
1	2	0

To find the optimal cut-off probability, we have to find the classification cut-off probability ' $p$ ' for which the total penalty [Eq. (11.37)] is minimum:

$$\underset{p}{\text{Min}}[2 \times P_{10}(p) + 1 \times P_{01}(p)] \quad (11.38)$$

Table 11.19 gives the total penalty cost for different classification cut-off probabilities. The minimum total cost is 0.756 and the corresponding cut-off probability is 0.25. Thus, the optimal classification cut-off probability is 0.25 when  $C_{10} = 2$  and  $C_{01} = 1$ .

**TABLE 11.19** Cost-based model for selection of classification cut-off probability

Classification Cut-off Probability ( $p$ )	$p_{11}$	$p_{10}$	$p_{00}$	$p_{01}$	Total Penalty
0.05	0.987	0.013	0.128	0.872	0.898
0.1	0.944	0.056	0.313	0.687	0.799
0.15	0.885	0.115	0.446	0.554	0.784
0.2	0.85	0.15	0.543	0.457	0.757
<b>0.25</b>	<b>0.816</b>	<b>0.184</b>	<b>0.612</b>	<b>0.388</b>	<b>0.756</b>
0.3	0.752	0.248	0.692	0.308	0.804
0.35	0.662	0.338	0.763	0.237	0.913
0.4	0.607	0.393	0.805	0.195	0.981
0.45	0.538	0.462	0.849	0.151	1.075
0.5	0.436	0.564	0.898	0.102	1.23
0.55	0.393	0.607	0.922	0.078	1.292
0.6	0.278	0.722	0.954	0.046	1.49
0.65	0.192	0.808	0.976	0.024	1.64
0.7	0.132	0.868	0.993	0.007	1.743
0.75	0.073	0.927	0.998	0.002	1.856
0.8	0.047	0.953	1	0	1.906
0.85	0.013	0.987	1	0	1.974

**TABLE 11.19** Cost-based model for selection of classification cut-off probability—Continued

Classification Cut-off Probability ( $p$ )	$p_{11}$	$p_{10}$	$p_{00}$	$p_{01}$	Total Penalty
0.9	0	1	1	0	2
0.95	0	1	1	0	2

## 11.10 | GAIN CHART AND LIFT CHART

Gain chart and lift chart are two measures that are used for measuring benefits of using the model and are used in business contexts such as target marketing. In target marketing or marketing campaigns, the customer responses to campaign are usually very low (in many cases the customers who respond to marketing campaigns are less than 1%). The organization will incur cost for each customer contact and hence would like to minimize the cost of marketing campaign and at the same time achieve desired response level from the customers. The gain and lift chart are obtained using the following steps:

1. Predict the probability  $Y = 1$  (positive) using LR model and arrange the observation in the decreasing order of predicted probability [i.e.,  $P(Y = 1)$ ].
2. Divide the data sets into deciles. Calculate the number of positives ( $Y = 1$ ) in each decile and cumulative number of positives up to a decile.
3. Gain is the ratio between cumulative number of the positive observations up to a decile to total number of positive observations in the data. Gain chart is a chart drawn between *gain* on vertical axis and *decile* on the horizontal axis.
4. Lift is the ratio of number of positive observations up to decile  $i$  using the model to the expected number of positives up to that decile  $i$  based on a random model. Lift chart is the chart between *lift* on the vertical axis and the corresponding *decile* on the horizontal axis.

$$\text{Gain} = \frac{\text{Cumulative number of positive observations upto decile } i}{\text{Total number of positive observations in the data}} \quad (11.39)$$

$$\text{Lift} = \frac{\text{Cumulative number of positive observations upto decile } i \text{ using LR model}}{\text{Cumulative number of positive observations up to decile } i \text{ based on random model}} \quad (11.40)$$

To illustrate the gain and lift chart, we will be using the data set ‘bank marketing data set’ available at the University of California, Irvine (UCI) machine learning repository. Data describes a problem in which a bank is interested in predicting which customers may respond to their direct marketing campaign to open a **term deposit** with the bank. The response variable  $Y = 1$  implies that the customer opens a term deposit after the campaign and 0 otherwise. The marketing campaign is based on the phone calls. The variables used for prediction of  $Y$  are listed in Table 11.20 (note that the data described in Table 11.20 is modified version of the original data in UCI website). The data set has a total of 4521 observations, out of which 521 customers subscribed term deposit (approximately 11.5%) and the remaining 4000 did not subscribe the term deposit (file name: Bank Marketing.xlsx).

**TABLE 11.20** Data dictionary<sup>2</sup>

S. No.	Variable	Variable Type	Code in SPSS Output
1	Term deposit subscription ( $Y$ )	Categorical	1 = Subscribed to term deposit 0 = Did not subscribe
2	Age	Numerical	Age
3	Marital Status	Categorical with 3 levels	1. Single 2. Married 3. Divorced
4	Education	Categorical with 4 levels	1. Primary 2. Secondary 3. Tertiary 4. Unknown
5	Job	Categorical with 5 levels	1. Blue Collar 2. Management 3. Self Employed 4. Unemployed 5. Others
6	Credit Default	Binary	1 = Had defaulted in the past 0 = No default in the past
7	Balance (average monthly bank balance)	Numerical	Balance
8	Housing Loan	Binary	1 = Customer has a housing loan 0 = Otherwise
9	Personal Loan	Binary	1 = Customer has a personal loan 0 = Otherwise
10	Current Campaign (Number of times a customer was contacted in the current campaign)	Numerical	Campaign
11	Previous Campaign (Number of times a customer was contacted in the previous campaign)	Numerical	Previous

The SPSS logistic regression output based on Forward LR (forward likelihood ratio) is shown in Table 11.21. Only steps 1, 2 and 8 are shown in Table 11.21.

<sup>2</sup> The example is prepared based on the sample data set downloaded from UCI repository. S. Moro, P. Cortez and P. Rita. *A Data-Driven Approach to Predict the Success of Bank Telemarketing*. Decision Support Systems, Elsevier, 62:22–31, June 2014.

**TABLE 11.21** Variables in the equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 <sup>a</sup>	Previous	0.143	0.021	46.931	1	0.000	1.154
	Constant	-2.139	0.050	1825.886	1	0.000	0.118
Step 2 <sup>b</sup>	Housing loan	-0.693	0.095	52.821	1	0.000	0.500
	Previous	0.152	0.021	51.934	1	0.000	1.165
	Constant	-1.797	0.065	772.696	1	0.000	0.166
	Age	0.013	0.004	8.940	1	0.003	1.014
	Blue collar	-0.323	0.145	4.937	1	0.026	0.724
	Married	-0.410	0.101	16.335	1	0.000	0.664
	Tertiary	0.207	0.105	3.868	1	0.049	1.229
	Housing loan	-0.574	0.099	33.464	1	0.000	0.563
	Personal loan	-0.704	0.167	17.736	1	0.000	0.495
	Campaign	-0.092	0.024	15.122	1	0.000	0.912
	Previous	0.143	0.021	45.918	1	0.000	1.154
	Constant	-1.872	0.219	72.935	1	0.000	0.154

The logistic regression model is given by

$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = -1.872 + 0.143 \text{ Previous} - 0.092 \text{ Campaign} - 0.704 \text{ Personal Loan} \\ - 0.574 \text{ Housing Loan} + 0.207 \text{ Tertiary} - 0.410 \text{ Married} - 0.323 \text{ Blue Collar} \\ + 0.013 \text{ Age}$$

Using the above equation, we can calculate the probability that a customer will subscribe a term deposit, that is,  $P(Y=1)$ . Since the data has 4521 cases, each decile will have approximately 452 observations. The gain using this model is shown in Table 11.22.

**TABLE 11.22** Gain and gain percentage using the model

Decile	Number of Observations	Number of Positives without Model	Number of Positives using Model	Cumulative Positives using the Model	Gain	Gain Percentage
1	452.1	52.1	223	223	0.4280	42.80%
2	904.2	104.2	122	345	0.6622	66.22%
3	1356.3	156.3	74	419	0.8042	80.42%
4	1808.4	208.4	38	457	0.8772	87.72%
5	2260.5	260.5	27	484	0.9290	92.90%
6	2712.6	312.6	11	495	0.9501	95.01%

(Continued)

**TABLE 11.22** Gain and gain percentage using the model—Continued

Decile	Number of Observations	Number of Positives without Model	Number of Positives using Model	Cumulative Positives using the Model	Gain	Gain Percentage
7	3164.7	364.7	18	513	0.9846	98.46%
8	3616.8	416.8	3	516	0.9904	99.04%
9	4068.9	468.9	4	520	0.9981	99.81%
10	4521	521	1	521	1.0000	100.00%

Gain can be interpreted as the gain in identifying customers who are likely to subscribe compared to a random model. Since the percentage of positives in the data is 521, by randomly targeting customers, we will expect approximately 52 customers to subscribe in each decile, whereas using the model and arranging them in descending probability of subscription, we observe that in the first decile there are 223 customers who will subscribe term deposit, which is 42.80% of all customers who have subscribed term deposit. Similarly, in the second decile 122 customers will subscribe term deposit against 52 from a random model. The cumulative number of customers who will subscribe term deposit in third decile is 419. Note that in a random model, the marketing team has to target 80% of the customers to ensure 80% of the subscribers are targeted, but using a logistic regression model, we need to target only 30% of the customers. Gain chart for the data described in the example is shown in Figure 11.9.

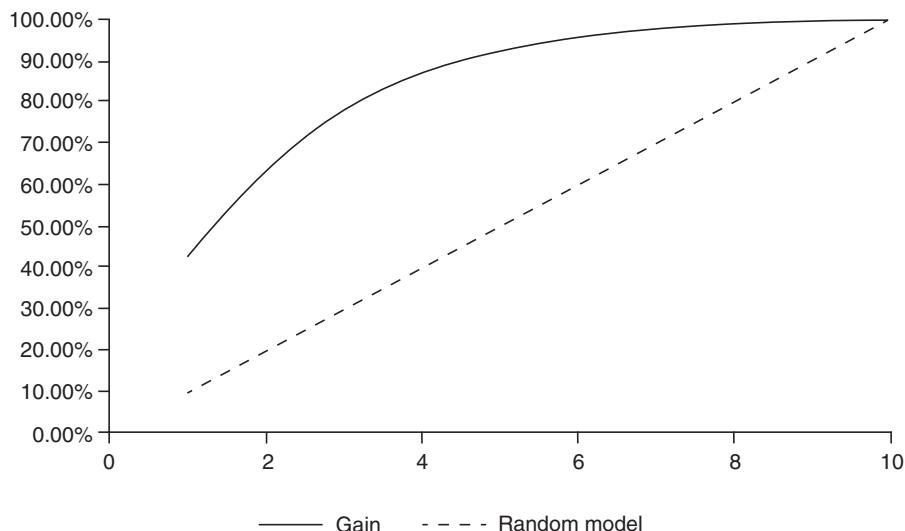
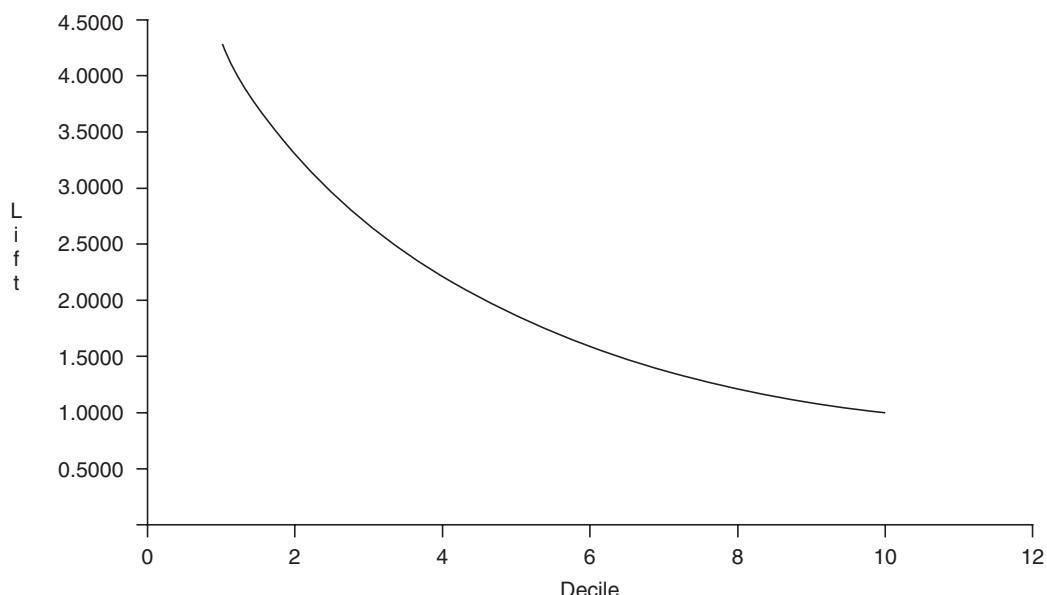
**FIGURE 11.9** Gain chart.

Table 11.23 shows the lift associated with the data. Lift is a measure of performance of the logistic regression model. For example, in Table 11.23, the response is 42.80% using LR model against 10% in a random model, thus the lift is 4.280. That is, targeting the customers using the model can capture 4.28 times the number of subscribers compared to a random model in decile 1.

**TABLE 11.23** Lift calculations

Decile	Cumulative Number of Positives without Model ( $A$ )	Number of Positives using Model	Cumulative Positives using Model ( $B$ )	Gain	Lift ( $B/A$ )
1	52.1	223	223	0.4280	4.2802
2	104.2	122	345	0.6622	3.3109
3	156.3	74	419	0.8042	2.6807
4	208.4	38	457	0.8772	2.1929
5	260.5	27	484	0.9290	1.8580
6	312.6	11	495	0.9501	1.5835
7	364.7	18	513	0.9846	1.4066
8	416.8	3	516	0.9904	1.2380
9	468.9	4	520	0.9981	1.1090
10	521	1	521	1.0000	1.0000

Lift chart is shown in Figure 11.10.

**FIGURE 11.10** Lift chart.

**SUMMARY**

- Classification problem is an important category of problems in analytics, and logistic regression is one of the most popular techniques used for solving classification problems.
- In a logistic regression model, the dependent variable  $Y$  takes finite discrete values. The regression parameters in logistic regression are estimated using maximum likelihood estimation.
- Mathematically, regression parameter in a logistic regression model captures change in the log ratio of odds for unit change in the independent variable value.
- Logistic regression is a generalized linear model (GLM) in which the residuals do not follow normal distribution.
- Logistic regression provides the probability of occurrence of the event; final class is usually predicted using a classification cut-off probability.
- The accuracy of an LR model is measured using metrics such as sensitivity, specificity,  $F$ -score, precision, and area under the ROC curve. Final model selection may be carried out using any of these metrics.
- The optimal classification cut-off probability is usually calculated using Youden's Index or cost-based approach.
- Many strategies such as forward likelihood ratio and forward Wald are used for selecting variables while building LR model.
- Gain and Lift charts are two approaches used while solving classification problems with imbalanced data sets.

**MULTIPLE CHOICE QUESTIONS**

- Which one of the following problems is not a classification problem?
  - Customer responding to an e-mail campaign.
  - Students placed (finding job) in campus placement.
  - Cash withdrawn from an ATM machine.
  - Predicting outcome of a football match.
- Deviance in a logistic regression model should be
  - Maximum.
  - Minimum.
  - Less than the chi-square critical value with  $df$  equal to number of variables added.
  - Greater the chi-square critical value with  $df$  equal to the number of variables added.
- In a logistic regression, the logit function is  $\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = -4 + 0.25X$ . Then the equation for logit function  $\ln\left(\frac{P(Y=0)}{1-P(Y=0)}\right)$  is
 

(a) $-0.25 + 4X$	(b) $0.25 - 4X$	(c) $4 - 0.25X$	(d) $4 + 0.25X$
------------------	-----------------	-----------------	-----------------
- In a logistic regression, the regression coefficient corresponding to a predictor variable is interpreted as
  - Change in  $P(Y=1)$  for unit change in the predictor variable value.
  - Change in odds for unit change in the predictor variable value.
  - Change in odds ratio for unit change in the predictor variable value.
  - Change in ln-odds ratio for unit change in the predictor variable value.
- In a logistic regression model, the statistical significance of a predictor variable is tested using
 

(a) Omnibus test.	(b) Wald's test.	(c) $F$ -test.	(d) $t$ -test.
-------------------	------------------	----------------	----------------

6. The area under the ROC curve (AUC) represents
  - (a) The maximum accuracy of the logistic regression model.
  - (b) Ratio of sensitivity to the specificity.
  - (c) Difference between sensitivity and specificity.
  - (d) Proportion of concordant pairs in the data set.
7. Youden's Index provides the best classification cut-off, when
  - (a) Sensitivity and specificity are equally important.
  - (b) Sensitivity and precision are equally important.
  - (c) The number of positives in the data set is more than the number of negatives.
  - (d) The number of negatives in the data set is more than the number of positives.
8. If in a data set with 250 positives, an LR model classifies 200 positives correctly, the specificity is
  - (a) 0.8
  - (b) 0.2
  - (c) 1.25
  - (d) Can't say
9. A logistic regression model between a customer churn as dependent variable and income and age as independent variables is developed. The corresponding LR model is

$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = 1.75 - 0.25 \text{Age} + 0.005 \text{Income}$$

where  $Y = 1$  implies churn. We can interpret that

- (a) The probability of churn increases as the age increases and decreases as income increases.
- (b) The probability of churn decreases as the age increases and increases as income increases.
- (c) Age has more impact on churn compared to income.
- (d) The probability of churn increases as age and income increases.

A binary logistic regression model is developed for probability of default in payment as dependent variable and the credit score as the predictor variable. The corresponding logit function is given by:

$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = 15 - 0.25 \times \text{credit Score} \quad (11.41)$$

$Y = 1$  indicates default. Answer Questions 10 to 12 using Eq. (11.41):

10. The value of credit score for which the probability of default in payment and probability of no default are equally likely is
  - (a) 575
  - (b) 600
  - (c) 450
  - (d) 500
11. The probability of default when the credits score = 300 is
  - (a) 0.9994
  - (b) 1
  - (c) 0.5
  - (d) 0.9770
12. A bank declines loan application when the probability of default is 0.2 (or 20%) or higher. The minimum credit score required to secure a loan from this bank is
  - (a) 675
  - (b) 650
  - (c) 625
  - (d) 666

## EXERCISES

1. A bank is interested in predicting which customers may respond to its direct marketing campaign to open a **term deposit** with the Bank. The response variable  $Y = 1$  implies that the customer opens a term deposit after the campaign (responds to the campaign) and 0 otherwise. The marketing campaign is based on the phone calls. The variables used for prediction of  $Y$  are listed in Table 11.20.

A logistic regression model is developed using 'campaign' as the explanatory variable. The SPSS outputs are shown in Tables 11.24–11.27.

$$\text{Model 1: } \ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 \times \text{Campaign}$$

**TABLE 11.24** Omnibus tests of model coefficients

		Chi-square	df	Sig.
Step 1	Step	22.524	1	
	Block	22.524	1	
	Model	22.524	1	

**TABLE 11.25** Model summary

Step	−2 Log Likelihood	Cox & Snell R-Square	Nagelkerke R-Square
1	3208.476	0.005	0.010

**TABLE 11.26** Classification table<sup>a</sup>

	Observed	Predicted		Percentage Correct	
		Subscription – Y			
		0	1		
Step 1	Subscription – Y	0	4000	0	
		1	521	0	
	Overall Percentage			88.5	

<sup>a</sup>The cut value is 0.500.

**TABLE 11.27** Variables in the equation

	B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 <sup>a</sup>	Campaign	−0.096	0.023	17.147	1	0.908
	Constant	−1.796	0.071	641.302	1	.000

<sup>a</sup>Variable(s) entered on step 1: campaign.

- (a) Using the information provided in Tables 11.24–11.27, calculate the value of  $-2LL_0$  ( $-2$  log likelihood function value when there is no variable in the model).
- (b) Comment whether model 1 is statistically significant.
- (c) Using model 1, comment on the impact of variable campaign on subscription of term deposit. Quantify the impact of variable "campaign" on term deposit subscription.

Model 2 is developed using 'job' as the predictor variable.

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \times \text{BlueCollar} + \beta_2 \times \text{Selfemployed} + \beta_3 \times \text{Management} + \beta_4 \times \text{Unemployed}$$

SPSS output for Model 2 is given in Tables 11.28 and 11.29.

**TABLE 11.28** Model summary

Step	–2 Log likelihood	Cox & Snell R-Square	Nagelkerke R-Square
1	3204.014 <sup>a</sup>	0.006	0.012

**TABLE 11.29** Variables in the equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 <sup>a</sup>	Blue Collar	−0.627	0.141	19.805	1	0.000	0.534
	Self Employed	−0.285	0.190	2.256	1	0.133	0.752
	Management	0.060	0.114	0.275	1	0.600	1.062
	Unemployed	−0.264	0.300	0.778	1	0.378	0.768
	Constant	−1.916	0.065	873.228	1	0.000	0.147

<sup>a</sup>Variable(s) entered on step 1: BlueCollar, SelfEmployed, Management, Unemployed.

- (d) Calculate the probability of term deposit subscription for the job category ‘others’.
- (e) Calculate the probability of subscription of term deposit for different job categories. Which job category has the highest probability of term deposit subscription? Use 5% significance level.

A stepwise regression model was developed using all explanatory variables. The SPSS outputs are shown in Tables 11.30 and 11.31. Eight independent variables are added to the LR model, Table 11.31 shows only the last step. Area under ROC curve is provided in Figure 11.11. The significance value is 0.05.

**TABLE 11.30** Classification Table<sup>a</sup>

	Observed	Predicted		Percentage Correct	
		Subscription – Y			
		0	1		
Step 1	Subscription – Y	0	3924	98.1	
		1	498	4.4	
	Overall Percentage			87.3	
Step 2	Subscription – Y	0	3883	97.1	
		1	472	9.4	
	Overall Percentage			87.0	
Step 3	Subscription – Y	0	3886	97.2	
		1	471	9.6	
	Overall Percentage			87.1	
Step 4	Subscription – Y	0	3871	96.8	
		1	452	13.2	
	Overall Percentage			87.1	

(Continued)

**TABLE 11.30** Classification Table<sup>a</sup>—Continued

		Observed	Predicted		Percentage Correct	
			Subscription – Y			
			0	1		
Step 5	Subscription – Y	0	3714	286	92.9	
	Subscription – Y	1	406	115	22.1	
	Overall Percentage				84.7	
Step 6	Subscription – Y	0	3726	274	93.2	
	Subscription – Y	1	399	122	23.4	
	Overall Percentage				85.1	
Step 7	Subscription – Y	0	3748	252	93.7	
	Subscription – Y	1	397	124	23.8	
	Overall Percentage				85.6	
Step 8	Subscription – Y	0	3721	279	93.0	
	Subscription – Y	1	393	128	24.6	
	Overall Percentage				85.1	

<sup>a</sup>The cut value is 0.200.

**TABLE 11.31** Variables in the equation (last iteration of the model)

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 8	BlueCollar	-0.323	0.145	4.937	1	0.026	0.724
	Married	-0.410	0.101	16.335	1	0.000	0.664
	Age	0.013	0.004	8.940	1	0.003	1.014
	Tertiary	0.207	0.105	3.868	1	0.049	1.229
	Housing Loan	-0.574	0.099	33.464	1	0.000	0.563
	Personal Loan	-0.704	0.167	17.736	1	0.000	0.495
	Campaign	-0.092	0.024	15.122	1	0.000	0.912
	Previous	0.143	0.021	45.918	1	0.000	1.154
	Constant	-1.872	0.219	72.935	1	0.000	0.154

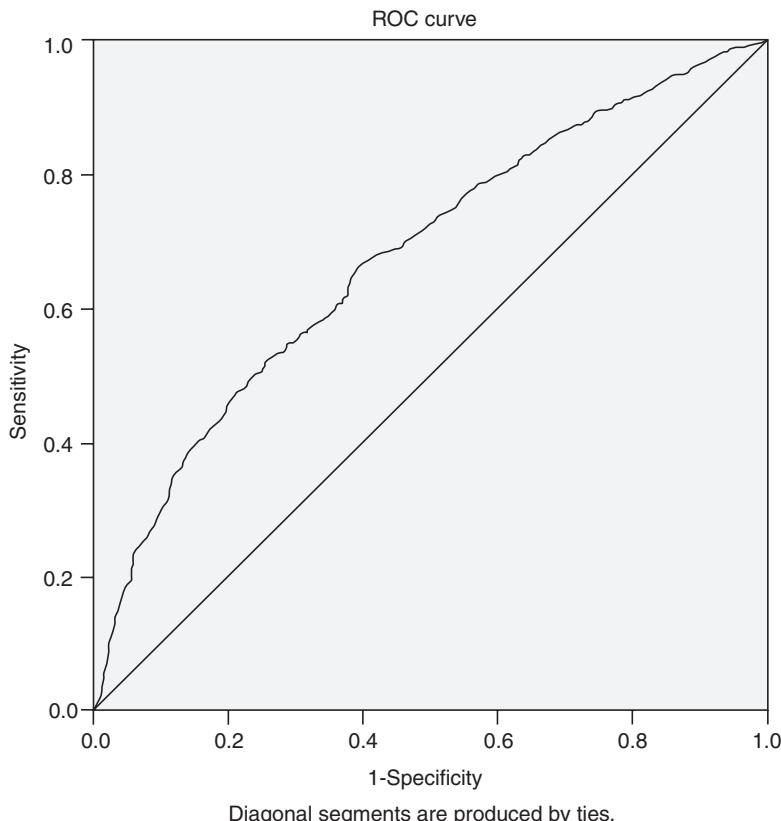


FIGURE 11.11 ROC Curve.

TABLE 11.32 Area under the curve

Test Result Variable(s): Predicted probability				
Area	Std. Error <sup>a</sup>	Asymptotic Sig. <sup>b</sup>	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.678	0.013	0.000	0.653	0.704

The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased. <sup>a</sup>Under the nonparametric assumption. <sup>b</sup>Null hypothesis: true area = 0.5.

Use the stepwise logistic regression model to answer to Questions (f)–(h).

- (f) Consider the variable, ‘marital status’. Customers with which ‘marital status’ are more likely to respond to the marketing campaign? Clearly state the reasons.
- (g) For two randomly selected customers, one who subscribed the term deposit and another who has not subscribed, what is the probability that probability of subscription is greater than the probability of no subscription?
- (h) Calculate the optimal classification cut-off probability using Youden’s Index.

2. Box office success of Bollywood movies was analysed using the following variables using logistic regression model. The data description is provided in Table 11.33.

**TABLE 11.33** Data description

S. No.	Variable	Variable Type	Code in SPSS output
1	Box office success ( $\gamma$ )	Categorical	1 = Success 0 = Failure
2	Release Data	Categorical with 4 levels	1 = Festival Season (FS) 2 = Holiday Season (HS) 3 = Long Weekend (LW) 4 = Other Season (OS)
3	Genre	Categorical with 5 levels	1 = Action (Action) 2 = Drama (Drama) 3 = Romance (Romance) 4 = Comedy (Comedy) 5 = Others (Other-G)
4	Movie Content	Categorical with 3 levels	Masala (Masala) Sequel (Sequel) Others (Other_C)
5	Director Category	Categorical with 3 levels	Director_A Director_B Director_O
6	Lead Actor Category	Categorical with 3 levels	Actor_A Actor_B Actor_O
7	Item Song	Binary variable	1 (Movie has an item song) 0 (otherwise)
8	Budget	Numerical (in crores of rupees)	Budget
9	YouTube Views	Numerical	YouTube-V
10	YouTube Likes	Numerical	YouTube-L
11	YouTube Dislikes	Numerical	YouTube-D

A logistic regression model was developed using Budget as independent variable and box office success as the dependent variable  $\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \times \text{Budget}$ . The SPSS model output is shown in Tables 11.34–11.36.

**TABLE 11.34** Omnibus tests of model coefficients

		Chi-square	df	Sig.
Step 1	Step	4.000	1	0.046
	Block	4.000	1	0.046
	Model	4.000	1	0.046

**TABLE 11.35** Classification Table<sup>a</sup>

		Predicted			Percentage Correct	
		Success Failure		0		
Observed		1	0			
Step 1	Success	1	2	17	10.5	
	Failure	0	3	41	93.2	
Overall Percentage					68.3	

<sup>a</sup>The cut value is 0.500.

**TABLE 11.36** Variables in the equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 <sup>a</sup>	Budget	-0.016	0.008	3.825	1	0.050	0.984
	Constant	1.621	0.503	10.395	1	0.001	5.058

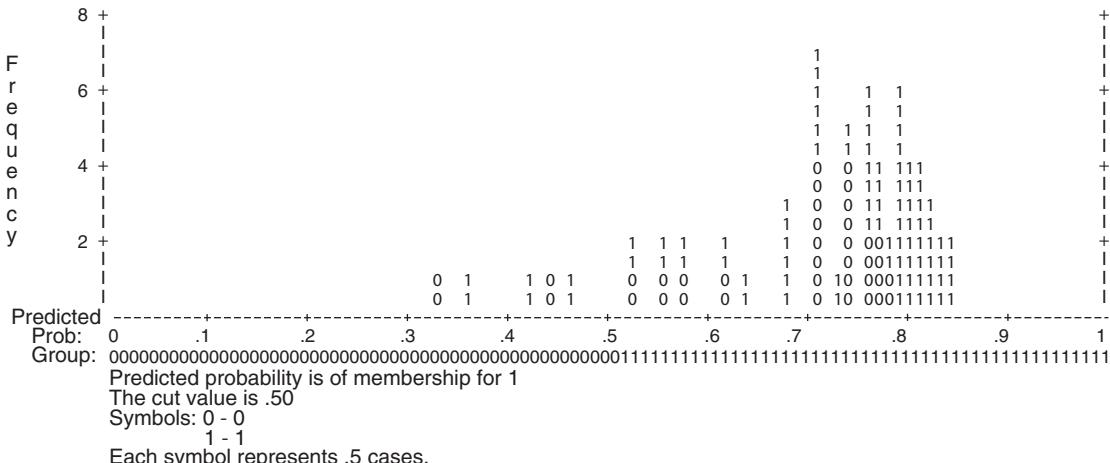
<sup>a</sup>Variable(s) entered on step 1: Budget.

Answer questions (a)–(d) based on Tables 11.34–11.36 and classification plot shown in Figure 11.12.

- (a) Calculate the budget for which the box office success and failure are equally likely.
  - (b) Is there a sufficient evidence to conclude that the higher budget movies are more likely to fail at the box office?
  - (c) A production house is making a movie with 100 crore budget, what is the success probability for this movie?
  - (d) Calculate the approximate sensitivity, specificity and precision for the classification cut-off probability of 0.6 using the classification plot in Figure 11.12.

## Step number: 1

### Observed groups and predicted probabilities



**FIGURE 11.12** Classification plot for model 1.

A second model is developed using the variable, ‘item song’. The SPSS output is shown in Tables 11.37 and 11.38.

**TABLE 11.37** Classification Table<sup>a</sup>

	Observed	Predicted		Percentage Correct	
		Success Failure			
		0	1		
Step 1	Failure	0	11	8	57.9
	Success	1	20	24	54.5
	Overall Percentage				55.6

<sup>a</sup>The cut value is 0.700.

**TABLE 11.38** Variables in the equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 <sup>a</sup>	Item Song	-0.501	0.202	6.151	1	0.013	0.606
	Constant	1.099	0.408	7.242	1	0.007	3.000

<sup>a</sup>Variable(s) entered on step 1: Item Song.

- (e) Calculate the difference in success probabilities for movies with item song and movies without item song.
  - (f) Which is a better model – budget as an independent variable versus item song as an independent variable?. Clearly state your reasons.
  - (g) Develop a logistic regression model using all significant variable in the data provided in Bollywood Box Office Success.xlsx. Calculate AUC, what can you infer from the value of AUC.
3. An insurance company is planning to develop a target marketing strategy for selling health insurance packages to its potential customers. The cost of contacting a new customer is approximately INR 150. The profit earned on average per annum per insurance holder is INR 180. The company has collected data on 8800 past customer contacts. The data dictionary is provided in Table 11.39.

**TABLE 11.39** Insurance customer contact data

S. No.	Variable	Variable Type	Code in SPSS Output
1	Insurance (Y)	Categorical	1 = Purchased insurance 0 = Did not buy insurance
2	Health	Categorical	1 = Healthy 0 = Not healthy
3	Age	Numerical	Numerical value measured in years
4	Any Limitation	Categorical	1 – Has limitation (person with disability) 0 – Has No limitation
5	Gender	Categorical	1 = Male 0 = Female

**TABLE 11.39** Insurance customer contact data—Continued

S. No.	Variable	Variable Type	Code in SPSS Output
6	Education	Categorical (4 levels)	deg_nd (Under Graduate) deg_hs (High School) deg_ged (Graduation) deg_others (Other qualifications)
7	Marital Status	Categorical	1 – Married 0 – Unmarried
8	Employment Type	Categorical	1 = Selfemp (Self employed) 0 = Others
9	Family Size	Numerical	Familysz (total size of the family)
10	Region	Categorical (4 categories)	reg_ne (North East) reg_mw (Mid-West) reg_so (South) reg_w (West)
11	Race	Categorical (3 categories)	race_AA (African American) race_white (White) race_others (Others)

The company is interested in developing a logistic regression model to understand purchase of insurance across different regions. Table 11.40 shows the logistic regression SPSS output for various regions as explanatory variables.

**TABLE 11.40** Variables in the equation

	B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 <sup>a</sup>	reg_ne	.479	.083	33.170	1	.000
	reg_mw	.645	.081	62.884	1	.000
	reg_so	.153	.068	5.153	1	.023
	Constant	1.118	.052	469.570	1	.000

<sup>a</sup>Variable(s) entered on step 1: reg\_ne, reg\_mw, reg\_so.

Use Table 11.40 to answer the following questions:

- What is the probability that a customer from mid-west will buy insurance if contacted by the company?
- If the company contacts 4 potential customers in Mid-west, what is the probability that at least 3 of them will buy insurance?

A second model is developed using family size, region, race and interaction between family size and race\_AA (code = FSIZEAA). The SPSS output for the new model is shown in Tables 11.41 and 11.42.

**TABLE 11.41** Classification Table<sup>a</sup>

	Observed	Predicted		Percentage Correct	
		Insured			
		0	1		
Step 1	Insured	0	0	1750 .0	
		1	0	7050 100.0	
Overall Percentage				80.1	

<sup>a</sup>The cut value is .500.

**TABLE 11.42** Variables in the equation

	<i>B</i>	S.E.	Wald	<i>df</i>	Sig.	Exp( <i>B</i> )	95% C.I. for EXP( <i>B</i> )	
							Lower	Upper
Step 1 <sup>a</sup>	familysz	-0.123	0.018	47.649	1	0.000	0.884	0.853 0.915
	reg_ne(1)	-0.318	0.079	16.276	1	0.000	0.728	0.623 0.849
	reg_mw(1)	-0.452	0.077	34.262	1	0.000	0.636	0.547 0.740
	reg_we(1)	0.163	0.070	5.446	1	0.020	1.177	1.026 1.350
	FSIZEAA	0.106	0.047	5.180	1	0.023	1.112	1.015 1.218
	race_AA	-0.470	0.209	5.051	1	0.025	0.625	0.415 0.942
	race_wht	0.161	0.128	1.569	1	0.210	1.174	0.913 1.510
	Constant	2.169	0.189	131.295	1	0.000	8.749	

<sup>a</sup>Variable(s) entered on step 1: familysz, reg\_ne, reg\_mw, reg\_we,FSIZEAA, race\_AA, race\_wht.

Use Tables 11.41 and 11.42 to answer to the following questions:

- (c) Identify the variables that increase the probability of buying an insurance (as the value variable increases) and the variables that decrease the probability of buying the insurance as the value of the variable increases (consider only variables that are significant at  $\alpha = 0.05$ ).
- (d) Is there a statistically significant impact of race on insurance purchase?
- (e) Interpret the co-efficient value of 0.106 for the interaction variable FSIZEAA (familysz x race\_AA).
- (f) Calculate the sensitivity, precision and the specificity of the model when the cut-off probability is 0.5.

## Case Study

### HR Analytics at ScaleneWorks – Behavioural Modelling to Predict Reneging<sup>3</sup>

ScaleneWorks supports several information technology (IT) companies in India with their talent acquisition. In 2015, Indian IT industry directly employed 3.5 million workers.<sup>4</sup> Acquiring new talent is always a challenging and time-consuming task, especially in IT since the hired person has to handle fast changing technology. In many cases, it is difficult to find the exact match for the job specified. If an offer is denied, then the Human Resource (HR) department has to repeat the entire recruitment process resulting in additional effort from the top management.

On 30 April 2014, Sanjay Shelvankar, Co-founder and the Chief Executive Officer of ScaleneWorks, had called for the meeting of its top management, which included Ashish Tiwari, Vice President and Head of Operations and Sharon George, Head of Technology and Strategy. Sanjay wanted to brainstorm with Ashish and Sharon on recent challenges the company was facing as talent management consultants. Sanjay began the meeting with the following statement:

We all know that talent acquisition is becoming a tough task, consuming time and effort of HR department and top management of many organizations. Even after they find the right talent, there is no guarantee that the person will join the organization if an offer is made.

Sharon George, one of the co-founders of the company, said:

If we can predict in advance whether someone will accept an offer or not, it will help companies to reduce their talent acquisition effort significantly. I think, many companies are now using analytics to address HR problems, I saw a video in YouTube on how Google is using analytics for promotions.

Ashish Tiwari, the other co-founder of the company, concurred and all of them decided to explore the possibility of using analytics to improve the talent acquisition process. ScaleneWorks had collected data from its past talent acquisition drives with key information such as current cost-to-company (CTC), expected CTC, offered CTC, locations and so on. All the founders of the company were convinced that it is possible to develop an early warning system that can help the companies to predict in advance on whether or not a person will accept an offer.

### **ScaleneWorks**

ScaleneWorks People Solutions LLP (ScaleneWorks) is a Bangalore based start-up that commenced its operations in the summer of 2010. ScaleneWorks was conceived by a team of HR practitioners

<sup>3</sup> Copyright © The Indian Institute of Management Bangalore (IIMB). The case is authored by Rahul Kumar and Professor U Dinesh Kumar and is distributed through Harvard Business Publishing as part of IIMB's case collection. Reproduced with permission from IIM Bangalore. The case is not intended to serve as an endorsement, source of primary data or effective handling of decision or business process.

<sup>4</sup> Source: Economic Times, February 17, 2015. Available at [http://articles.economictimes.indiatimes.com/2015-02-27/news/59585092\\_1\\_economic-survey-e-commerce-market-export-market](http://articles.economictimes.indiatimes.com/2015-02-27/news/59585092_1_economic-survey-e-commerce-market-export-market)

**Continued...**

comprising Sanjay Shelvankar, Ashish Tiwari and Sharon George who had already scripted successful corporate careers and were from three different areas of expertise such as, Technology Consulting, Talent Acquisition and Marketing. Their combined vision was to build an organization of great value and to position it amongst the most respected Talent Acquisition Solutions provider globally within the next 5 years; this was reflected in the way they carefully chose their customers and engaged with them. ScaleneWorks sees itself as the first true end-to-end Talent Acquisition Solutions organization which has the passion to bring together decades of experience in Technology Consulting and Talent Acquisition areas to usher in a paradigm shift in the way Talent Acquisition is practiced in today's ultra-demanding business environment. Scalene Works not only advises its customers on where their Talent Acquisition practices are, but also recommends and implements individually tailored, viable solutions using analytics.

Business process re-engineering with its three tenets – People Capability, Process Maturity and Technology Adoption – form the core ability of the company to provide customers with an enterprise-class customized solution to address their Talent Acquisition challenges. They bring in deep domain knowledge of how Talent Acquisition happens in corporates and provide viable recommendations to their customers.

### Current Business Challenge

Client service is all about the quality of the people involved in delivering business. However, one of the major challenges for Sanjay and his clients revolved around managing a quality workforce. Organizations spend tremendous amount of time and energy to create a homogenous environment where people thrive and succeed. Despite all the effort to keep an environment that is conducive, people leave organizations in search of better opportunities. In order to fill the vacuum, HR is bound to recruit new talent, thus forming a vicious circle in between attrition and recruitment; and in order to mitigate this, organizations keep trying to bridge the gap by strengthening their recruitment processes and creating a culture of inclusivity.

Sanjay wanted to find a unique solution which goes beyond the process aspect of human resource management. At first, Sanjay identified and prioritized the renege problem and put forward in a subtle way:

“In my opinion, a significant proportion of the candidates do not join the company that has made an offer. If we can identify them in advance, then companies don't have to waste their resources.”

Although this problem is generic, for a case-in-study we've identified a particular client of Scalene Works. According to Sanjay in a typical IT services company, the number of people not joining the company varies anywhere between 15% and 35% of all the people who accepted the offer.

### Case Study Continued...

Sanjay went ahead to explain the impact of this problem from time, cost and quality perspective. The impact may seem minimal if the number of offers rolled out to candidates revolves around hundreds in a year. But if the offers rolled out surpasses the thousands mark, the magnitude of impact rises exponentially.

He elaborated the impact for a client<sup>5</sup> where 12,000 offers are rolled out every year. At 30% renege rate, about 3600 candidates would accept the offer and then not join the company. Even with the most conservative estimates, on an average organizations would have spent 15 hours in the recruitment lifecycle, effectively indicating a humongous loss of 54,000 man hours wasted from an organization's perspective by one client alone. This involves the time spent by the business in interviewing the candidate whose value is more than the mere numbers of hours.

Renege has greater impact on the cost of talent acquisition. The entire recruitment lifecycle, starting from sourcing resumes till candidate is deemed fit for recruitment, involves various agencies. These agencies work in tandem with the talent team to screen for candidates who would fit the profile. There is a payout which happens to these agencies for their involvement in the recruitment cycle. It is estimated that if one accounts for cost of recruitment associated with the renege candidates, we would find that the cost of hiring goes up anywhere between 10% and 15% according Scalene Works.

Ashish elaborated a scenario of business impact of renege:

If a candidate sends in a mail rejecting the offer just 10 days before his date of joining and if the business has already committed to the client and had made an entire plan of on-boarding the new joinee to the project, then what do we do. Either we go and tell this to the client and make a miserable situation which no one would like to do in front of the customer, or, we look for an alternative. Most of us look for alternative and fill up the position so as to make business go as usual. But in doing so, we cannot expect to get the same quality of resource as was the case with the one who reneged.

So Sanjay and Ashish wanted to know answers to the following questions:

1. What are the key drivers that influence the candidate joining/not-joining a company?
2. What rules can be used to predict the acceptance or rejection of the offer?
3. How to devise a predictive algorithm to calculate the probability of acceptance of an offer and joining the company after offer acceptance stage?

Scalene Works had captured several data related to the applicants. The variable description is provided in **Exhibit 1**. **Exhibit 2** describes the recruitment process in detail.

<sup>5</sup> Name of the client is not revealed to ensure confidentiality.

**Continued...**

**EXHIBIT 1** A total of 12000 records spanning a year was captured. The variables were as shown in the Table

S. No.	Variable Name	Variable Description
1	Candidate reference number	Unique number to identify the candidate
2	DOJ extended	Binary variable identifying whether candidate asked for Date of joining extension (Yes/No)
3	Duration to accept the offer	Number of days taken by the candidate to accept the offer (Scale variable)
4	Notice period	Notice period to be served in the parting company before candidate can join this company (Scale variable)
5	Offered band	Band offered to the candidate based on experience, performance in interview rounds (C0/C1/C2/C3/C4/C5/C6)
6	Percentage hike expected	Percentage hike expected by the candidate (Scale variable)
7	Percentage hike offered	Percentage hike offered by the company (Scale variable)
8	Joining bonus	Binary variable indicating if joining bonus was given or not (Yes/no)
9	Gender	Gender of the candidate (Male/Female)
10	Candidate source	Source from which resume of the candidate was obtained (Employee referral/Agency/ Direct)
11	REX (in Yrs.)	Relevant years of experience of the candidate for the position offered (Scale variable)
12	LOB	Line of business for which offer was rolled out (Categorical variable)
13	Date of Birth	Date of birth of the candidate
14	Joining location	Company location for which offer was rolled out for candidate to join (Categorical variable)
15	Candidate relocation status	Binary variable indicating whether candidate has to relocate from one city to another city for joining (Yes/No)
16	HR Status	Final joining status of candidate (Joined/Not Joined)

Source: Scaleneworks.

## EXHIBIT 2

### The Recruitment Process

The recruitment process at ScaleneWorks follows the usual Talent acquisition lifecycle (TALC):

Sourcing → Screening → Selection → Fitment & Offer → Post offer follow-up (PoFu)

The recruitment process for a company starts when the Resource Management Group (RMG) performs a yearly demand planning in conjunction with business units, sales team and Talent Acquisition Group (TAG). TAG finally takes over the recruitment process to meet the finalized demand pipeline.

**Continued...**

**Sourcing** involves looking for resumes which can fit in the different schemas of demand. Typical channels to source resume involves:

1. Job portals
2. Employee referral
3. Advertisement/Walk-ins
4. Direct
5. Vendors/Consultancy
6. Internal database of sources resumes
7. Social networking sites

Sourcing from vendors/consultancy is most expensive while sourcing through internal database and social networking sites falls under least expensive way of sourcing resumes. However, companies prefer to go with more than one way of sourcing so as to balance the cost, quality and effort required to get an optimal mix of resumes. Management and TAG looks into the conversion rate for each channel and cost per channel to arrive at the channel mix to be used to meet the resource demand pipeline.

**Screening** can broadly be divided into two types:

1. Hygiene screening involves scanning the resume for notice period to be served by the candidate, gap in education, previous companies of employment, etc.
2. Technical screening involves matching the skills mentioned in the resume with the desirable skills mentioned in the job description for a particular position.

The screening process is a time-consuming and strenuous process. A team of HR executives are involved in screening nearly 12000–15000 resumes for every client every month. The screening process also comes with a pre-determined service level agreement (SLA) with the client. This SLA puts a cap on the number of resumes which should not be rejected in the selection round. From a client's perspective, this SLA is needed to ensure the quality of resume being screened by the HR executives.

**Selection** process involves multiple rounds of interview for the candidates whose resumes have been screened and cleared. Typically, selection process would have:

1. *Technical assessment 1*: This primarily would be a telephonic round.
2. *Technical assessment 2*: This may be a second telephonic round or a face to face interview.
3. *Final round*: This would be a face to face interview round with senior management.
4. *HR round*: This round of interview is aimed at understanding the communication and interpersonal skills of the candidate.
5. *Customer round*: This round's primary objective is to ensure the comfort level of the candidate and gain client's confidence in the skills of the candidate.

**Continued...**

Initially, the customer round may or may not take place for all the candidates. The objective of selection process is to evaluate the candidate on technical/functional skills, process and tools knowledge, domain knowledge and behavioural aspects.

**Fitment and offer** is a function of the score given by the interview panel in different rounds of selection process. Demonstration of skills, knowledge and attitude as judged by the interview panel finally gives a fitment calculation score. The offer roll-out or rejection is guided by this score.

If the candidate is deemed fit for the position, a final offer is rolled out which details each and every aspect of the employment. There is also an online system through which offer is rolled out to the candidates, wherein they can either accept or reject the offer online.

**Post offer follow up (PoFu)** process involves the HR executive to be in touch with the candidate to whom the offer has been rolled out in order to ensure that the candidate joins the company post completion of the notice period being served with the parting company. The operations in this team are carried out by sub-teams organized as follows:

1. **Document collection team:** If the candidate accepts the offer, within 24 hours a link is sent to the candidate through which the candidate has to upload all the documents which aid in completing the joining formality. The target for the document management team is to get all the relevant documents uploaded and verified for completeness within five days. Once the documents are uploaded and verified for completeness, the document management team intimates the third-party vendor for background verification process. The background verification needs to be completed within 15–20 days of candidates having submitted the documents.
2. **Advanced PoFu team:** In case the document collection team finds candidates unwillingness to submit the document, the case gets escalated to the advanced PoFu team. This team works closely with the clients TAG to sort out the issue and make amendments in the offer, if needed.
3. **Reneging management team:** In case the candidate accepts the offer and has uploaded the document successfully, this team takes the wagon forward to ensure that the candidate joins the company post completion of the notice period and on date of joining agreed by the candidate. This team keeps following up with the candidate using a structured six-stage process. Every stage has a series of questionnaire which the team uses to rate the probability of candidate joining the company. The objective of this team is to provide 70% predictability by stage 2. The objective for the stages is detailed below.

Stages	Objectives to Check
Stage 1 (SLA: 0–3 days)	Offer letter received Offer accepted Joining form filled Documents uploaded in the system Resignation submitted check

**Continued...**

Stages	Objectives to Check
Stage 2 (SLA: Stage 1 + 10%*Notice period)	Last working day confirmation
Stage 3 (SLA: Stage 1 + 20%*Notice period)	Replacement found
Stage 4 (SLA: Stage 1 + 40%*Notice period)	Knowledge transition completed
Stage 5 (SLA: Stage 1 + 60%*Notice period)	Resignation acceptance mail Relieved from current role
Stage 6 (SLA: Stage 1 + 80%*Notice period)	Induction details received Date of joining confirmed Joined or not joined

Duration between calls is subject to change. Stage-wise objectives, issues, queries/unresolved queries, incoming calls from candidate, holidays, BGV, recruiter requests, etc. may impact the duration. Finally, renege cases are escalated to the advanced PoFu team for discussion, if needed.

**EXHIBIT 3** The six stage questionnaire. In general, the table below depicts the dialect process which follows for the six stages

Stage 1			
Greeting	Mandatory	Resigned Check	Mandatory
Introduction	Mandatory	Resignation Acceptance Check	Mandatory
Call Recipient Check	Mandatory	Notice Period	Mandatory
Conversation Starter	Preferable	Replacement and KT	Mandatory
Setting The Agenda	Mandatory	Retention Efforts	Preferable
Job Offer Check	Mandatory	Tentative/Confirmed LWD	Mandatory
Congratulations on OA	Mandatory	Referral Pitch	Preferable
Reason for accepting the offer	Preferable	Tentative/Confirmed DOJ	Mandatory
Prelude	Preferable	Relocating Yes/No	Preferable
Documents Uploaded Check	Mandatory	Accommodation Required Yes/No	Preferable
Joining Form Filled Check	Mandatory	Induction Details	Mandatory
BGV Status	If Required	Preparation For Closing the Call	Mandatory
Current Company Details	Mandatory	Call Closing	Mandatory
Company Reactions	If Required	Prediction	Mandatory
Resignation Process	Mandatory		
Stage 2		Stage 3	
Greeting	Mandatory	Greeting	Mandatory
Introduction	Mandatory	Introduction	Mandatory
Call Recipient Check	Mandatory	Call Recipient Check	Mandatory

(Continued)

**Continued...**

**EXHIBIT 3** The six stage questionnaire. In general, the table below depicts the dialect process which follows for the six stages—Continued

Conversation Starter	Preferable	Conversation Starter	Preferable
Setting The Agenda	Preferable	Setting The Agenda	Preferable
Documents Uploaded Check	If Required	Documents Uploaded Check	If Required
Joining Form Filled Check	If Required	Joining Form Filled Check	If Required
BGV Status	If Required	BGV Status	If Required
Company Reactions	If Required	Company Reactions	If Required
Resignation Process	Mandatory	Resignation Process	Mandatory
Notice Period	Mandatory	Notice Period	Mandatory
Replacement and KT	Mandatory	Replacement and KT	Mandatory
Retention Efforts	Preferable	Retention Efforts	Preferable
Confirmed LWD	Mandatory	Confirmed LWD	Mandatory
Confirmed DOJ	Mandatory	Confirmed DOJ	Mandatory
Referral Pitch	If Required	Referral Pitch	If Required
Preparation For Closing the Call	Mandatory	Preparation For Closing the Call	Mandatory
Call Closing	Mandatory	Call Closing	Mandatory
Prediction	Mandatory		
<b>Stage 4</b>		<b>Stage 5</b>	
Greeting	Mandatory	Greeting	Mandatory
Introduction	Mandatory	Introduction	Mandatory
Call Recipient Check	Mandatory	Call Recipient Check	Mandatory
Conversation Starter	Preferable	Conversation Starter	Preferable
Setting The Agenda	Preferable	Setting The Agenda	Preferable
Documents Uploaded Check	If Required	Documents Uploaded Check	If Required
Joining Form Filled Check	If Required	Joining Form Filled Check	If Required
BGV Status	If Required	BGV Status	If Required
Company Reactions	If Required	Company Reactions	If Required
Resignation Process	Mandatory	Resignation Process	Mandatory
Notice Period	Mandatory	Notice Period	Mandatory
Replacement and KT	Mandatory	Replacement and KT	Mandatory
Retention Efforts	Preferable	Retention Efforts	Preferable
Confirmed LWD	Mandatory	Confirmed LWD	Mandatory
Confirmed DOJ	Mandatory	Confirmed DOJ	Mandatory
Referral Pitch	If Required	Referral Pitch	If Required
Relocation Planned Yes/No	Preferable	Relocation Planned Yes/No	Preferable

Continued...

**EXHIBIT 3** The six stage questionnaire. In general, the table below depicts the dialect process which follows for the six stages—Continued

Accommodation Planned Yes/No	Preferable	Accommodation Planned Yes/No	Preferable
Preparation For Closing the Call	Mandatory	Induction Details	Mandatory
Call Closing	Mandatory	Preparation For Closing the Call	Mandatory
		Call Closing	Mandatory
<b>Stage 6</b>			
Greeting	Mandatory		
Introduction	Mandatory		
Call Recipient Check	Mandatory		
Conversation Starter	Preferable		
Setting The Agenda	Preferable		
Documents Uploaded Check	If Required		
Joining Form Filled Check	If Required		
BGV Status	If Required		
Company Reactions	If Required		
Resignation Process	Mandatory		
Notice Period	Mandatory		
Replacement and KT	Mandatory		
Retention Efforts	Preferable		
Confirmed LWD	Mandatory		
Confirmed DOJ	Mandatory		
Referral Pitch	If Required		
Relocation Planned Yes/No	Preferable		
Accommodation Planned Yes/No	Preferable		
Induction Details	Mandatory		
Joined/ Not Joined Check	If Required		
Preparation For Closing the Call	Mandatory		
Call Closing	Mandatory		

Source: SclaneWorks.

Continued...

### CASE QUESTIONS

#### (use data set HR Analytics.xlsx)

1. Develop a logistic regression model that can be used by ScaleneWorks for predicting candidates who are unlikely to join after accepting the offer. Which are the variables having statistical significance on renege?
2. Devise a logistic regression function to calculate the probability of acceptance of an offer and finally joining the company after offer acceptance.
3. How would you interpret sensitivity, specificity, and model accuracy? Calculate the AUC, comment on the LR model developed using AUC.
4. What cut-off probability ScaleneWorks should use to classify joining and not joining the firm after accepting the offer?
5. How should one handle outliers in the data in case of a logistic regression model?
6. How should be the model deployment strategy for ScaleneWorks?

### REFERENCES

1. Cox D R and Snell E J (1989), "The Analysis of Binary Data (2nd Edition)", Chapman and Hall, London.
2. DeMaris A (1995), "A Tutorial on Logistic Regression", *Journal of Marriage and Family*, **57**(4), 956–968.
3. Feynman R (1988), "What do you Care What Other People Think", W W Norton & Company, New York.
4. Gastwirth J L (1971), "A General Definition of Lorenz Curve", *Econometrica*, **39**(6), 1037–1039.
5. Hosmer D W, Jovanovic B and Lemeshow S (1989), "Best Subsets Logistic Regression", *Biometrics*, **45**(4), 1265–1270.
6. Hosmer D W and Lemeshow S (2000), "Applied Logistic Regression, 2<sup>nd</sup> Edition", John Wiley and Sons, New York.
7. Kleinbaum D G and Klein M (2011). "Logistic Regression – A Self-Learning Text", Springer, New Delhi
8. Lawless J F and Singhal K (1987), "ISMOD-An all Subsets Regression Program for Generalized Linear Models I: Statistical and Computational Background", *Computer Methods and Programs in Biomedicine*, **24**, 117–124.
9. Nagelkerke N (1991), "A Note on General Definition of the Coefficient of Determination", *Biometrika*, **78**, 691–692.
10. Smith R J (1986), "Inquiry Faults Shuttle Management", *Science – New Series*, **232**(4757), 1488–1489.
11. Wilks S S (1938), "The Large Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses", *Annals of Mathematical Statistics*, **9**, 60–62.
12. Youden W J (1950), "Index for Rating Diagnostic Tests", *Cancer*, **3**, 32–35.

# 12

# Decision Trees

“Nothing is particularly hard if you divide it into small jobs”.

— Henry Ford

## LEARNING OBJECTIVES

- LO 12-1** Understand foundations of decision trees (also known as decision tree learning and classification trees) and how they are used for solving classification problems.
- LO 12-2** Learn how to construct tree techniques such as Chi-square Automatic Interaction Detection (CHAID) and Classification and Regression Tree (CART) and their application in solving classification problems.
- LO 12-3** Understand different splitting strategies such as Chi-square Test, F-test, Gini Impurity Index and entropy used in decision trees.
- LO 12-4** Understand how to generate business rules using decision trees.
- LO 12-5** Learn ensemble methods and the concept of random forest.
- LO 12-6** Understand concepts such as bagging and boosting.

## ESSENCE OF DECISION TREES

Decision Trees are collection of divide-conquer problem-solving strategies that use tree-like structure to predict the outcome of a variable. The tree starts with the root node consisting of the complete data and thereafter uses intelligent strategies to split the nodes (parent node) into multiple branches (thus creating children nodes). The original data is divided into subsets. This is done in order to create more homogenous groups at the children nodes. It is one of the most powerful predictive analytics techniques used for generating business rules.

### 12.1 | DECISION TREES: INTRODUCTION

Decision trees (also known as decision tree learning or classification trees) are a collection of predictive analytics techniques that use tree-like graphs for predicting the value of a response variable (or target variable) based on the values of explanatory variables (or predictors). It is one of the supervised learning algorithms used for predicting both the discrete and the continuous dependent variable. In a decision tree learning, when the response variable takes discrete values then the decision trees are called classification trees.

Decision trees are effective for solving classification problems in which the response variable (target variable) takes discrete values. Decision trees employ divide-and-conquer strategy in which the original

data is divided into multiple groups or subsets, and the strategy is to establish groups such that within groups the data is homogeneous. This means that the data in the groups is dominated by one class. Decision trees use the following criteria to develop the tree:

1. **Splitting Criteria:** Splitting criteria are used to split a node (set of data) into subsets.
2. **Merging Criteria:** When the predictor variable is categorical with  $n$  categories, it is possible that not all  $n$  categories may be statistically significant. Thus, few categories may be merged to create a compound or aggregate category.
3. **Stopping Criteria:** Stopping criteria is used for pruning the tree (stopping the tree from further branching) to reduce the complexity associated with business rules generated from the tree. Usually levels (depth) from root node (where each level corresponds to adding a predictor variable), minimum number of observation in a node for splitting are used as stopping criteria.

The following steps are used for generating decision trees:

1. Start with the **root node** in which all the data is present.
2. Decide on a splitting criterion and stopping criteria: The root node is then split into two or more subsets leading to tree branches (called edges) using the splitting criterion. Nodes thus created are known as **internal nodes**. Each internal node has exactly one incoming edge.
3. Further divide each internal node until no further splitting is possible or the stopping criterion is met. The **terminal nodes** (aka **leaf nodes**) will not have any outgoing edges.
4. Terminal nodes are used for generating business rules.
5. **Tree pruning** (a process for restricting the size of the tree) is used to avoid large trees and overfitting the data. Tree pruning is achieved through different stopping criteria.

There are many decision tree techniques and they differ in the strategy that they use for splitting the nodes. In this chapter two of the most popular decision tree techniques, namely, Chi-square Automatic Interaction Detection (CHAID) and Classification and Regression Trees (CART) will be discussed.

## 12.2 | CHI-SQUARE AUTOMATIC INTERACTION DETECTION (CHAID)

CHAID (Kass, 1980) is an extension of Automatic Interaction Detection (AID), which is designed to categorize the dependent variable using categorical predictors. The technique partitions the data set into mutually exclusive and exhaustive subsets using independent variables that result in hierarchical splitting of the original data thus resembling a tree-like structure. CHAID trees use statistical significance of independent variables to split the subset of the data (represented by nodes of the tree).

Initial models of CHAID tree were developed for discrete or categorical dependent variable and used Chi-square Test of Independence to split the data (and thus the name CHAID). Initial CHAID models have now been extended to partition numerical and ordinal dependent variables as well.

Depending on the nature of the dependent variable, the following statistical tests are used for splitting the nodes:

1. Chi-square Test of Independence when the response variable,  $Y$ , is discrete.
2.  $F$ -test when the response variable,  $Y$ , is continuous.
3. Likelihood Ratio Test when the response variable,  $Y$ , is ordinal.

The following steps are used in developing a CHAID tree:

1. Start with the complete training data in the **root node**.
2. Check the statistical significance of each independent variable depending on the type of dependent variable (Chi-square Test of Independence if the dependent variable is categorical,  $F$ -test if the dependent variable is continuous and Likelihood Ratio Test if the dependent variable is ordinal).
3. The variable with the least  $p$ -value, based on the statistical tests described in step 2, is used for splitting the data set thereby creating subsets (represented by the **internal nodes** in the tree). Bonferroni Correction is used for adjusting the significance level  $\alpha$ . Non-significant categories in a categorical predictor variable with more than two categories may be merged.
4. Using independent variables, repeat step 3 for each of the subsets of the data (that is, for each **internal node**) until:
  - (a) All the dependent variables are exhausted or they are not statistically significant at  $\alpha$ .
  - (b) The stopping criteria is met.
5. Generate business rules for the **terminal nodes** (nodes without any branches) of the tree.

### 12.2.1 | CHAID Tree Development

We will be using a sample of 800 observations (file name: German Credit Raing.Xlsx) from the German Credit Data (can be downloaded from the University of California, Irvine machine learning data repository<sup>1</sup>) to illustrate CHAID tree development. The data description is provided in Table 11.13. At each stage of the tree portioning, an appropriate hypothesis test has to be conducted for the variable selection.

In the case of German credit data, the dependent variable  $Y$  is categorical ( $Y = 0$  indicates good credit and  $Y = 1$  indicates bad credit) and thus we will have to use Chi-square Test of Independence to check whether there is a dependency relationship between the dependent variable and the independent variables. The null and alternative hypotheses are given below:

$$\begin{aligned} H_0: & \text{The variables } Y \text{ and } X \text{ are independent} \\ H_A: & \text{The variables } Y \text{ and } X \text{ are dependent} \end{aligned}$$

The contingency table between the target variable ( $Y$ ) and one of the predictor variables (checking account balance = 0 Deutsche Mark labelled as 0 DM) is shown in Table 12.1 along with expected frequencies.

---

<sup>1</sup> Source: <https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29>

**TABLE 12.1** Contingency table for German credit rating sample data and expected frequencies

Checking Account Balance	Credit Rating (Observed Frequencies)		Total	Credit Rating (Expected Frequencies)	
	Y = 1	Y = 0		Y = 1	Y = 0
0 DM = 1	99	110	209	62.44	146.56
0 DM = 0	140	451	591	176.56	414.44
Total	239	561	800	239	561

In Table 12.1, 0 DM = 1 implies balance in checking account = 0 DM, 0 DM = 0 implies the balance is other than 0 DM. The dependent variable Y = 0 implies good credit and Y = 1 implies bad credit. The expected frequencies are calculated using the following equation:

$$E_{ij} = \frac{i^{\text{th}} \text{ row sum} \times j^{\text{th}} \text{ column sum}}{\text{Total Sum}}$$

The statistic for Chi-square Test of Independence is given by

$$\chi^2 = \sum_{i=1}^n \sum_{j=1}^m \left( \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right) \quad (12.1)$$

The null and alternative hypotheses in this context are as follows:

$H_0$ : Credit classification, Y, and checking account balance 0 DM are independent

$H_1$ : Credit classification, Y, and checking account balance 0 DM are dependent

Using the values in contingency table (Table 12.1), the chi-square statistic is given by

$$\chi^2 = \frac{(99 - 62.44)^2}{62.44} + \frac{(110 - 146.56)^2}{146.56} + \frac{(140 - 176.56)^2}{176.56} + \frac{(451 - 414.44)^2}{414.44} = 41.322$$

The chi-square critical value is 3.841. Since the chi-square statistic is much greater than chi-square critical value, we reject the null hypothesis. The corresponding  $p$ -value is  $1.29 \times 10^{-10}$ . That is, there is a statistically significant relationship between credit classification (Y) and checking account balance 0 DM.

Using chi-square test of independence, we established that the credit classification and checking account balance 0 DM are dependent. When there are more than one variable, the model will select the variable with least  $p$ -value (if it is less than the significance value  $\alpha$ ) and split the node using that variable.

The CHAID tree split for the sample German credit data using checking account balance as the predictor variable is shown in Figure 12.1.

In Figure 12.1, the node 0 is the **root node**, which has 800 observations (561 good credits and 239 bad credits) and, since there are no outgoing branches from nodes 1 and 2, nodes 1 and 2 are

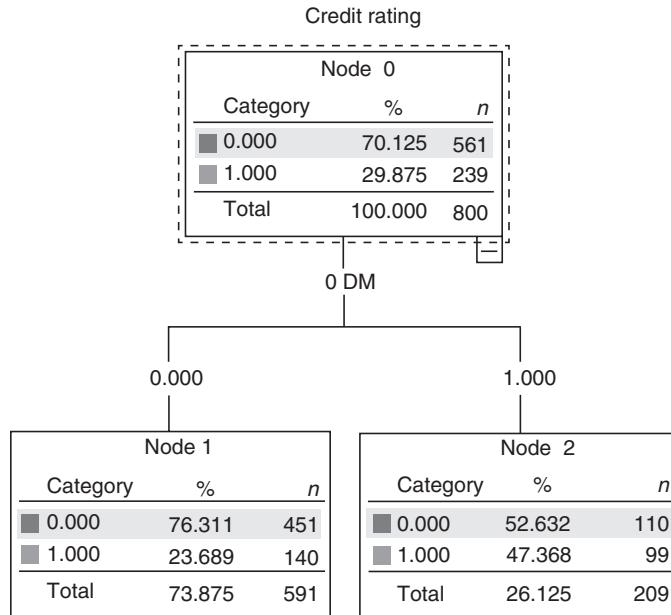


FIGURE 12.1 CHAIID tree German credit data.

**terminal nodes (leaf nodes).** Node 1 in Figure 12.1 corresponds to the value of the predictor variable value  $0 \text{ DM} = 0$  (that is, the checking account balance is other than  $0 \text{ DM}$ ) and node 2 in Figure 12.1 corresponds to the predictor variable value  $0 \text{ DM} = 1$  (that is, the checking account balance is  $0 \text{ DM}$ ).

The CHAIID tree in Figure 12.2 is created by incorporating all categories within the checking account balance. In Figure 12.2, node 1 corresponds to checking account balance to either  $0 \text{ DM}$  or less than  $200 \text{ DM}$ . The model has merged two checking account balance categories ( $0 \text{ DM}$  and less than  $200 \text{ DM}$ ) to create node 1. Node 2 represents the customers who do not have a checking account and node 3 represents those customers with more than  $200 \text{ DM}$  in their checking account balance.

## 12.2.2 | Bonferroni Correction

Since the method can result in several splits of the original data using many independent variables at multiple levels, hence while developing CHAIID tree, we test multiple hypotheses simultaneously. When multiple hypotheses are tested simultaneously, the significance value  $\alpha$  must be corrected to ensure the validity of all hypotheses simultaneously at  $\alpha = 0.05$ , that is, the Type I error remains as  $0.05$ .

In a hypothesis testing, the probability of rejecting a null hypothesis when it is true is Type I error ( $= \alpha$ ). However, when more than one hypotheses tests are conducted simultaneously, the Type I error increases. For example, assume that two hypotheses tests are conducted simultaneously at  $\alpha = 0.05$  in a CHAIID tree development. That is, a terminal node is a result of two hypotheses tests at  $\alpha = 0.05$ . Then the probability of retaining a null hypothesis when it is true is  $0.95$ . The probability of retaining both null hypotheses simultaneously when they are true is  $0.95 \times 0.95 = 0.9025$ . That is, the Type I

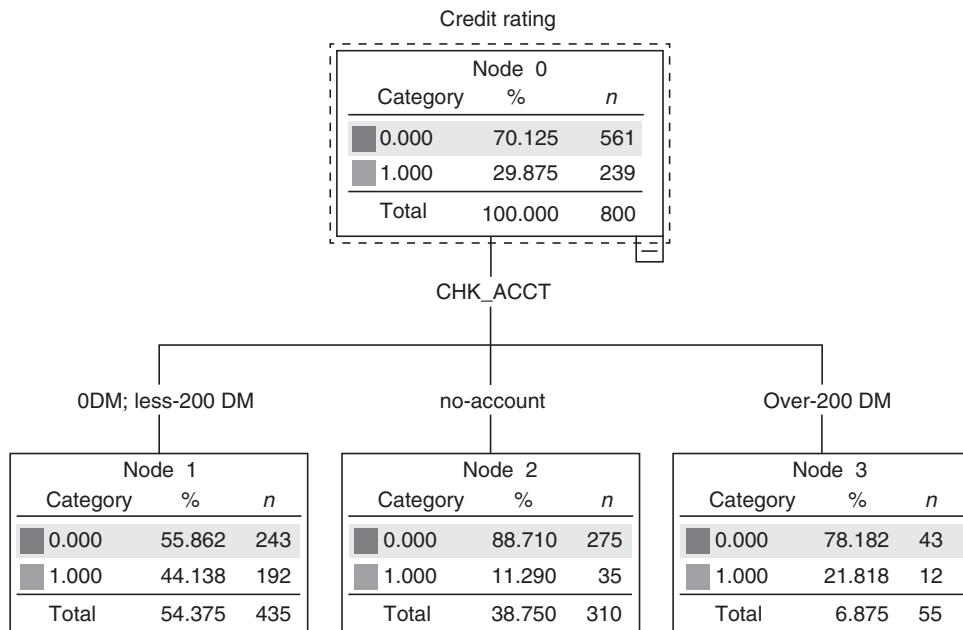


FIGURE 12.2 CHAID Tree with all categories within checking account balance.

error in this case is  $1 - 0.9027 = 0.0975$ . So, the Type I error has increased from 0.05 to 0.0975. The Type I error will further increase when we increase the number of simultaneous tests. For example, assume that we are testing 10 different hypotheses simultaneously at  $\alpha = 0.05$ . The probability of at least one significant result out of 10 tests is

$$\begin{aligned}
 P[\text{At least one significant result out of 10 tests}] &= 1 - P[\text{no significant result}] \\
 &= 1 - (1 - 0.05)^{10} = 0.4012
 \end{aligned}$$

That is, there is a 40% chance of making Type I error when a terminal node in CHAID tree is constructed by testing 10 different hypotheses at  $\alpha = 0.05$ .

The Bonferroni Correction sets the significant cut-off for individual test at  $\alpha/n$  instead of  $\alpha$  when  $n$  hypothesis tests are conducted simultaneously (Armstrong, 2014). That is, we set a lower Type I error ( $\alpha/n$ ) at individual tests, but this may increase the Type II error which is one of the criticisms of the Bonferroni Correction.

### 12.2.3 | Generating Business Rules using CHAID Tree

One of the advantages of decision trees is their use in generating business rules that can be deployed for decision-making. Figure 12.3 is a CHAID tree developed by using predictor variables such as checking account balance, marital status, and employment (in the German credit rating data) with a stopping criterion of 2 levels from root node.

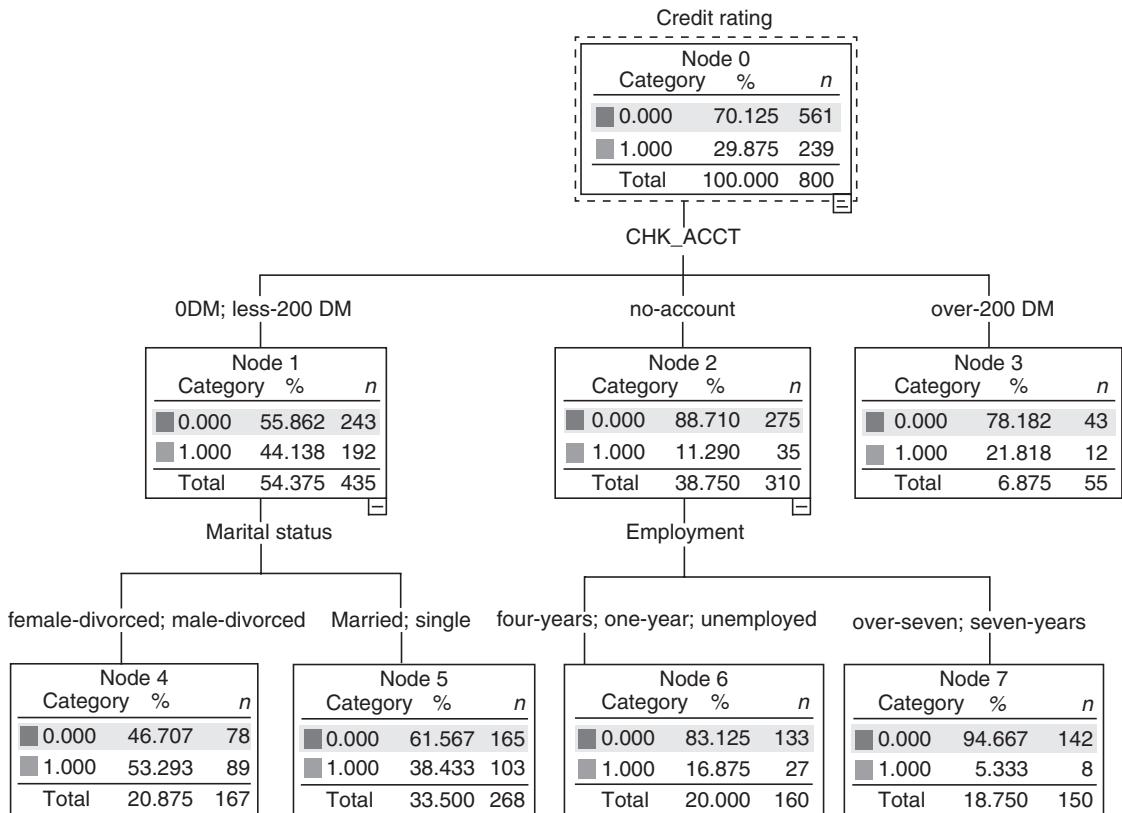


FIGURE 12.3 CHAID tree for German credit rating sample data.

In Figure 12.3, nodes 3 to 7 are **leaf nodes (terminal nodes)** since there are no other branches emerging out of these nodes. Each **terminal node** is classified based on the largest class in that node. For example, in node 3, the proportion of  $Y = 0$  (good credit) is 78.182% and proportion of bad credit ( $Y = 1$ ) is 21.818%. Since the majority class in node 3 is 0, all the data in this node will be classified as 0 using the business rule checking account balance over 200 DM. That is, when checking account balance is over 200 DM then the value of the outcome variable  $Y$  is 0.

Table 12.2 provides the list of business rules corresponding to the **leaf nodes** and the corresponding support (proportion of the data in that node).

TABLE 12.2 Business rules and support

Node	Business Rule	Support
3	Checking account balance is more than 200 DM, classify the outcome as $Y = 0$ . Classification accuracy is 78.18%.	6.875%
4	Checking account balance is either 0 DM or less than 200 DM <b>AND</b> the marital status is male divorced or female divorced, classify outcome as $Y = 1$ . Classification accuracy is 53.293%.	20.875%

(Continued)

**TABLE 12.2** Business rules and support—Continued

Node	Business Rule	Support
5	Checking account balance is either 0 DM or less than 200 DM <b>AND</b> the marital status is married or single, classify outcome as $Y = 0$ . Classification accuracy is 61.56%.	33.500%
6	No checking account <b>AND</b> the employment is 1. Unemployed or 2. One year or 3. Four years  Then classify the outcome as $Y = 0$ . Classification accuracy is 83.125%.	20.00%
7	No checking account <b>AND</b> the employment is either seven years or over seven years then classify the outcome as $Y = 0$ . The classification accuracy is 94.667%.	18.750%

For practical application, one may look for the business rules that have high accuracy as well as high support. For example, the accuracy of node 7 is 94.667%, and 18.75% of data (150 out of 800 observations) fall under this rule. Thus the business rule corresponding to node 7 is very effective. On the other hand, the rule corresponding to node 4 has a low accuracy of 53.293%.

## 12.3 | CLASSIFICATION AND REGRESSION TREE

Classification and Regression Tree (CART) is a common terminology that is used for a **Classification Tree** (used when the dependent variable is discrete) and a **Regression Tree** (used when the dependent variable is continuous).

Classification tree uses various impurity measures such as the **Gini Impurity Index** and **Entropy** to split the nodes. Regression Tree, on the other hand, splits the node that minimizes the **Sum of Squared Errors** (SSE).

CART is a binary tree wherein every node is split into only two branches. CHAID, however, can have more than two splits from a node. The following steps are used to generate a classification and a regression tree (Breiman *et al.*, 1984):

1. Start with the complete training data in the **root node**.
2. Decide on the **measure of impurity** (usually Gini impurity index or Entropy). Choose a predictor variable that minimizes the impurity when the parent node is split into **children nodes** [see Eq. (12.4)]. This happens when the original data is divided into two subsets using a predictor variable such that it results in the maximum reduction in the impurity in the case of discrete dependent variable or the maximum reduction in SSE in the case of a continuous dependent variable.
3. Repeat step 2 for each subset of the data (for each **internal node**) using the independent variables until:
  - (a) All the dependent variables are exhausted.
  - (b) The stopping criteria are met. Few stopping criteria used are number of levels of tree from the root node, minimum number of observations in parent/child node (e.g. 10% of the training data), and minimum reduction in impurity index.
4. Generate business rules for the **leaf (terminal) nodes** of the tree.

### 12.3.1 | Gini Impurity Index

Gini impurity index is one of the measures of impurity that is used by classification trees to split the nodes.

Assume that the original data has  $K$  classes (labelled as  $C_1, C_2, \dots, C_K$ ). Then the Gini impurity index at node  $t$  is given by (Breiman *et al.*, 1984)

$$GI(t) = \sum_{i=1}^K \sum_{j=1, j \neq i}^K P(C_i|t)P(C_j|t) = \sum_{i=1}^K P(C_i|t)(1 - P(C_i|t)) = 1 - \sum_{i=1}^K [P(C_i|t)]^2 \quad (12.2)$$

where

$GI(t)$  = Gini index at node  $t$

$P(C_i|t)$  = Proportion of observations belonging to class  $C_i$  in node  $t$

For a problem with two classes  $C_1$  and  $C_2$ , the Gini index is given by<sup>2</sup>

$$GI(t) = \sum_{i=1}^2 \sum_{j=1, j \neq i}^2 P(C_i|t)P(C_j|t) = 2 \times P(C_1|t)[1 - P(C_1|t)] = 1 - \sum_{i=1}^2 [P(C_i|t)]^2 \quad (12.3)$$

That is, if there are only two classes, then  $GI(t)$  is

$$GI(t) = 2 \times \text{Proportion of observations in class 1} \times \text{Proportion of observations in class 2}$$

A classification tree chooses the independent variable that minimizes the Gini impurity index. For a classification problem with 2 classes, the Gini index value will lie between 0 and 0.5. Higher value of the Gini index indicates higher impurity. At each step of the classification tree, the algorithm chooses the variable that provides the maximum reduction in Gini impurity. Consider Figure 12.4. The node  $t$  is split into left node and right node.

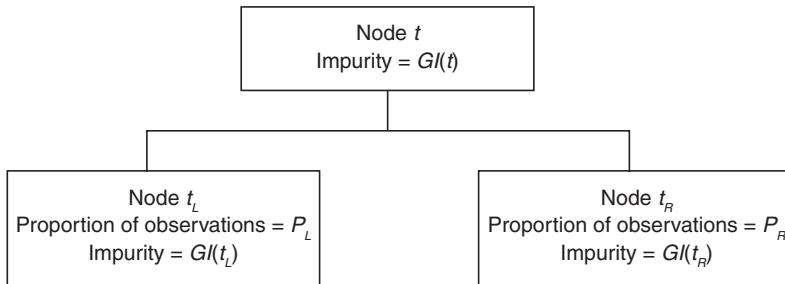


FIGURE 12.4 Splitting strategy in CART.

At node  $t$ , the algorithm will split it into two branches (labelled as node  $t_L$  and node  $t_R$ ) as shown in Figure 12.4. CART chooses a predictor variable,  $X_p$ , for splitting at node  $t$  that maximizes the function in Eq. (12.4) (Breiman *et al.*, 1984):

$$\underset{x_i}{\operatorname{Max}} [GI(t) - P_L GI(t_L) - P_R GI(t_R)] \quad (12.4)$$

<sup>2</sup> If there are two classes, and  $P_1$  is the proportion of data in class 1 and  $P_2$  is the proportion of data in class 2, the Gini index using Eq. (12.2) is  $2P_1P_2$ . Using  $(P_1 + P_2)^2 = P_1^2 + P_2^2 + 2P_1P_2$ , we get  $2 \times P_1P_2 = 1 - P_1^2 - P_2^2$ .

where

$$GI(t) = \text{Gini impurity at node } t$$

$$GI(t_L) = \text{Gini impurity at node } t_L$$

$$GI(t_R) = \text{Gini impurity at node } t_R$$

$$P_L = \text{Proportion of cases in node } t_L$$

$$P_R = \text{Proportion of cases in node } t_R$$

We will be discussing application of CART to predict success of a Bollywood movie at the box-office. Data description is provided in Table 12.3. Figure 12.5 shows a classification tree for the success/failure of a Bollywood movie ( $Y = 1$  implies success at the box office and  $Y = 0$  implies failure at the box office) based on the movie's release date (data set: Bollywood Data.Xlsx).

The predictor variable, the movie release date, has four categories:

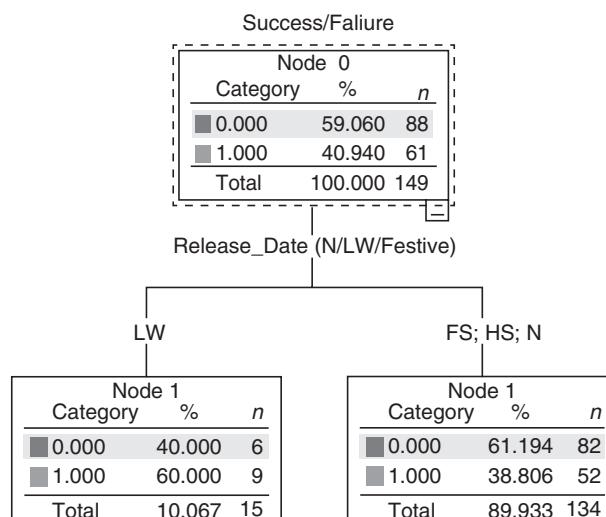
1. Long Weekend (LW)
2. Festive Season (FS)
3. Holiday Season (HS)
4. Normal (N)

At node 0, the Gini index is

$$GI(1) = 2 \times 0.59060 \times 0.40940 = 0.4835$$

Gini index for nodes 1 and 2 are 0.48 ( $= 2 \times 0.4 \times 0.6$ ) and 0.4749 ( $= 2 \times 0.61194 \times 0.38806$ ), respectively. The reduction in impurity is given by

$$[GI(t) - P_L GI(t_L) - P_R GI(t_R)] = [0.4835 - 0.10067 \times 0.48 - 0.89933 \times 0.4749] = 0.008$$



**FIGURE 12.5** Classification tree for success/failure of a Bollywood movie based on release date.

That is, when the predictor variable ‘release date’ is used for splitting the root node, the Gini impurity is reduced by 0.008. The decision maker can then set a minimum threshold  $\epsilon$  (say  $\epsilon = 0.0001$ ) as the minimum requirement in reduction in impurity for splitting a node.

When the data has more than one predictor variable, the model starts with a variable that gives the maximum reduction in impurity.

Figure 12.6 shows a classification tree developed using predictor variables as defined in Table 12.3 for the Bollywood movie data.

**TABLE 12.3** Predictor variables used for creating classification tree in Figure 12.6

Predictor Variable	Variable Type	Description
Budget	Continuous	Movie budget measured in crores of rupees
Production_House_Cat	Categorical	3 categories A, B and LK (lesser known)
Music_Dir_Cat	Categorical	3 categories A, B and LK (lesser known)
Lead_Actor_Cat	Categorical	3 categories A, B and LK (lesser known)
Director_Cat	Categorical	3 categories A, B and LK (lesser known)

A stopping criterion of 3 levels from the root node was used to develop the tree shown in Figure 12.6.

One can generate business rules based on the **leaf nodes** of the tree in Figure 12.6 (similar to Table 12.2). For example, from node 6 of the tree we understand that if the production house belongs to the category B or LK and if the budget is more than 6.5 crores, then there is more than 77% chance that the movie is likely to fail at the box office.

### 12.3.2 | Entropy

Entropy is another popular measure of impurity that is used in classification trees to split a node. Assume that there are  $K$  classes labelled  $C_1, C_2, \dots, C_K$ . The entropy at node  $t$  is given by

$$\text{Entropy}(t) = -\sum_{i=1}^K P(C_i | t) \times \log_2 P(C_i | t) \quad (12.5)$$

The value of entropy for node 0 of the tree in Figure 12.6 is  $0.9762 = (-0.59060 \log_2 0.59060 - 0.40940 \log_2 0.40940)$ . The value of entropy lies between 0 and 1, with a higher entropy indicating a higher impurity at the node. The value of entropy is higher than the value of Gini impurity index for a given proportion of positives and negatives. Figure 12.7 shows the plot of the Gini index and entropy for different values of  $P(C_i | t)$ . However, most commercial software tools use Gini impurity index for splitting the data.

## 12.4 | COST-BASED SPLITTING CRITERIA

Other than impurity measures such as Gini impurity index and entropy, decision makers may also use **Cost of Misclassification** to split the data.

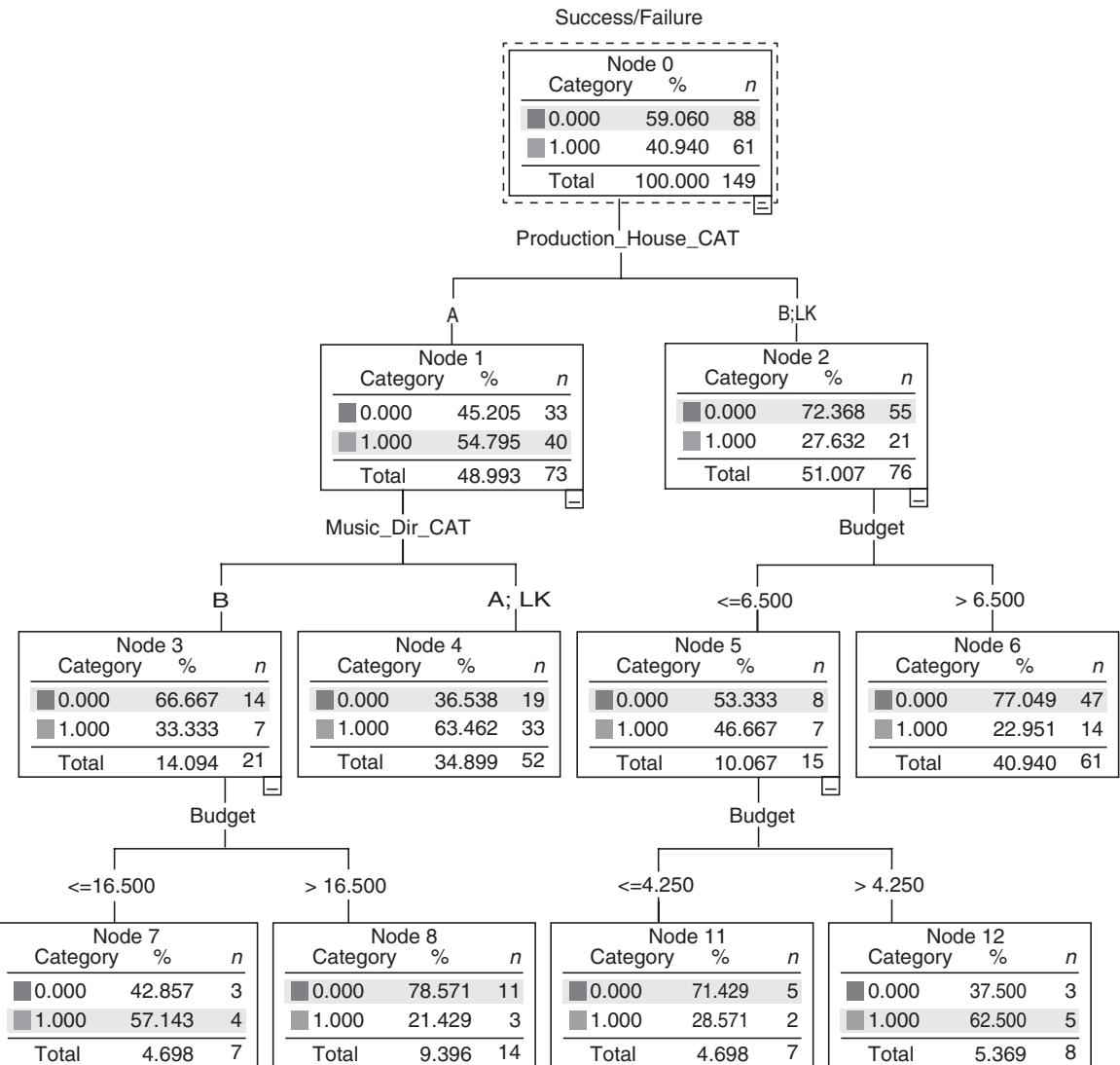
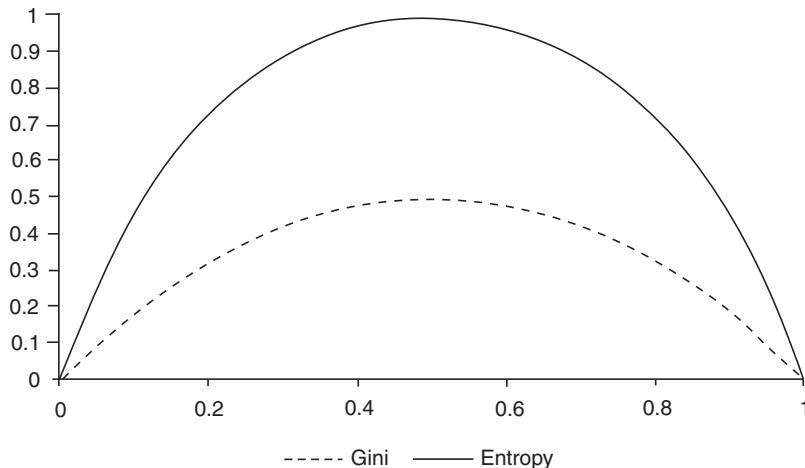


FIGURE 12.6 Classification tree for success/failure of a Bollywood movie.

Assume the penalty for misclassification as defined in Table 12.4.

TABLE 12.4 Misclassification cost

Observed	Predicted	
	0	1
0	0	$C_{01} (= 5)$
1	$C_{10} (= 1)$	0



**FIGURE 12.7** Comparison of Gini index and entropy.

The total penalty is  $C_{01}P_{01} + C_{10}P_{10}$ , where

$P_{01}$  = Proportion of 0 classified as 1

$P_{10}$  = Proportion of 1 classified as 0

$C_{01}$  = Penalty for classifying 0 as 1

$C_{10}$  = Penalty for classifying 1 as 0

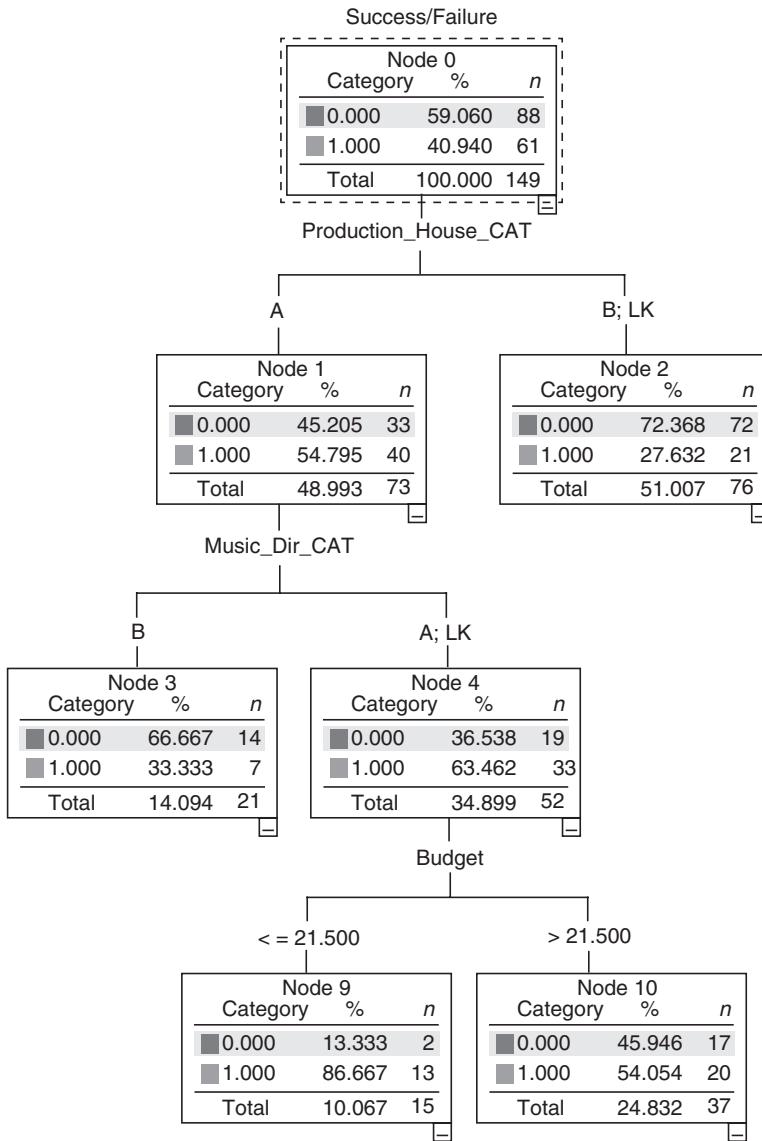
In Table 12.4, we assume that there is a penalty for misclassifying  $Y = 0$  (negative) as  $Y = 1$  and this is fixed at 5 ( $C_{10} = 5$ ). Similarly, the misclassification cost for misclassifying  $Y = 1$  (positive) as  $Y = 0$  (negative) is 1 ( $C_{01} = 1$ ).

Classification tree for the Bollywood data (success of the movie at the box office) using penalty cost minimization is shown in Figure 12.8. The tree in Figure 12.8 is different from the tree in Figure 12.6, which was developed using Gini impurity index from level 2 onwards. In many cases, the decision makers may prefer the cost-based classification compared to impurity measure-based classification.

## 12.5 | ENSEMBLE METHOD

The ensemble method is a machine-learning-algorithm that generates several classifiers (a classifier means a classification model) using different sampling strategies such as bootstrap aggregating. A majority-voting-approach may be used for classifying a new observation using the multiple classifiers that are developed. In the ensemble method, several techniques (such as logistic regression, CHAID, CART, etc.) are used. For a new observation, its class is identified using all the classifiers that are part of the ensemble method. Different classifiers are likely to classify a new observation into different categories. The final class of a new observation is decided based on a majority vote in which each classifier is given equal weightage.

Alternatively, different weights can be assigned to classifiers based on their accuracy (boosting algorithms such as **Adaptive Boosting** assign weight to each classifier based on its accuracy). The class of a new observation is then decided based on the value obtained using a linear combination of all the classifiers developed.



**FIGURE 12.8** Classification tree based on misclassification cost ( $C_{01}=5$  and  $C_{10}=1$ ).

## 12.6 | RANDOM FOREST

Random Forest is one of the popular ensemble method in which several trees (thus the name “forest”) are developed using different sampling strategies. One of the most frequently used sampling strategy is the **Bootstrap Aggregating** (or **Bagging**). Bagging is a random sampling with replacement. A new observation is classified by using all the trees developed in the random forest and majority voting is used for deciding the class. Random forests are developed using the following steps:

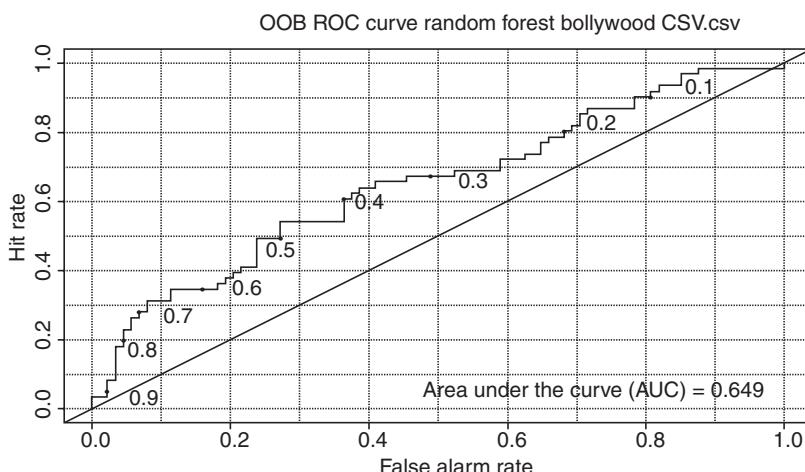
1. Assume that the training data has  $N$  observations. One needs to generate several samples of size  $M$  ( $M < N$ ) with replacement (called **Bagging**). Let the number of samples based on sampling of the training data set be  $S_1$ .
2. If the data has  $n$  predictors, sample  $m$  predictors ( $m < n$ ).
3. Develop trees for each of the samples generated in steps 1 using the sample of predictors from step 2 using CART.
4. Repeat step 3 for all the samples generated in step 1.
5. Predict the class of a new observation using majority voting based on all trees.

In general, random forest approach is expected to provide much higher accuracy compared to a single tree. However, one has to be aware of possible overfitting while using random forest. The model is validated using the validation data, known as **Out-of-Bag (OOB)** data. Consider a data  $(X, Y)$ ; this data may not be part of the training data set used for creating many trees in the random forest due to random sampling. All such cases that are not part of training data of a tree can be used as a validation data and such cases are called Out-Of-Bag data.

The random forest model using 500 trees and sampling 2 out of 5 variables in Table 12.3 is developed using **Rattle Package** (**Rattle** is one of the packages in open source software R) and the corresponding classification table is shown in Table 12.5. The corresponding **ROC curve** based on OOB data is shown in Figure 12.9. The overall accuracy of random forest is 64.42% and area under of ROC curve for the OOB data is 0.649.

**TABLE 12.5** Classification table based on random forest

Observed	Predicted		Accuracy
	0	1	
0	65	23	73.86%
1	30	31	50.81%
Overall Accuracy		64.42	



**FIGURE 12.9** OOB ROC curve (false alarm = 1 – specificity and hit rate = sensitivity).

**SUMMARY**

1. Decision trees are important set of techniques used in predictive analytics to solve problems associated with continuous as well as discrete dependent variables. However, decision trees are mostly used for solving classification problems and such trees are called classification trees.
2. Decision trees are supervised learning algorithms. They are developed using different splitting, merging, and stopping criteria.
3. There are several decision tree learning algorithms and they differ mainly in the splitting criteria.
4. Chi-square automatic interaction detection (CHAID) uses chi-square test of independence when the dependent variable is categorical, *F*-test when the dependent variable is continuous, and likelihood ratio test when the dependent variable is ordinal as splitting strategy.
5. In classification and regression tree (CART) impurity measures such as Gini impurity index or entropy are used as splitting criteria when the dependent variable is categorical and sum of squared errors (SSE) is used when the dependent variable is continuous.
6. Decision trees are an integral part of random forest algorithm. Random forest technique uses sampling with replacement (bagging) to create several trees and the class of a new observation is decided on the basis of majority voting.
7. One of the major advantages of decision tree learning algorithms is that it can be used for generating business rules from the model directly.

**MULTIPLE CHOICE QUESTIONS**

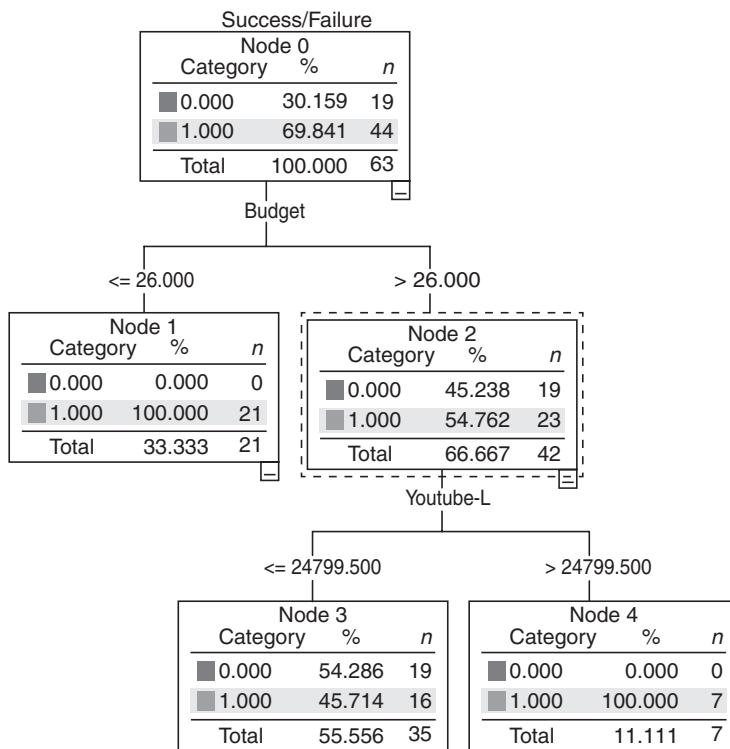
---

1. Decision tree such as CHAID and CART can be used only when the dependent variable is
  - (a) Discrete
  - (b) Continuous
  - (c) Interval
  - (d) All three (discrete, continuous and interval)
2. CHAID tree uses the following criteria for splitting:
  - (a) Chi-square Goodness of Fit Test for continuous predictor and *F*-test for discrete predictor
  - (b) Chi-square Test of Independence for all types of predictor variables
  - (c) Chi-square Test of Independence for discrete dependent variable and *F*-test for continuous dependent variable.
  - (d) *F*-test for all type of predictor variables
3. In a Classification and Regression Tree (CART), the splitting of node is based on
  - (a) Gini impurity index or entropy
  - (b) Impurity measures such as Gini index or entropy for classification tree and Sum of Squared Errors (SSE) for regression tree
  - (c) Sum of Squared Errors (SSE)
  - (d) F-test
4. The minimum and maximum values of Gini index are
  - (a) 0 and 0.5
  - (b) 0 and 1
  - (c) 0.5 and 1
  - (d) 1 and 2
5. For a classification problem with two classes, the proportion of positives at a node is 20%. The value of the Gini index at this node is
  - (a) 0.16
  - (b) 0.13
  - (c) 0.26
  - (d) 0.32
6. For a classification problem with two classes, the proportion of positives at a node is 80%. The value of the entropy at this node is
  - (a) 0.72
  - (b) 0.16
  - (c) 0.32
  - (d) 0.50

7. For a classification problem with two classes
- Gini index value is greater than or equal to entropy
  - Gini index value is less than or equal to entropy
  - Gini index value is greater than or equal to entropy when proportion of classes are equal
  - Cannot say
8. Bonferroni correction in CHAID is used to
- Split a data into subsets
  - Merge variables that are not statistically significant
  - Adjust significance value  $\alpha$  to correct Type I error
  - Prune the tree

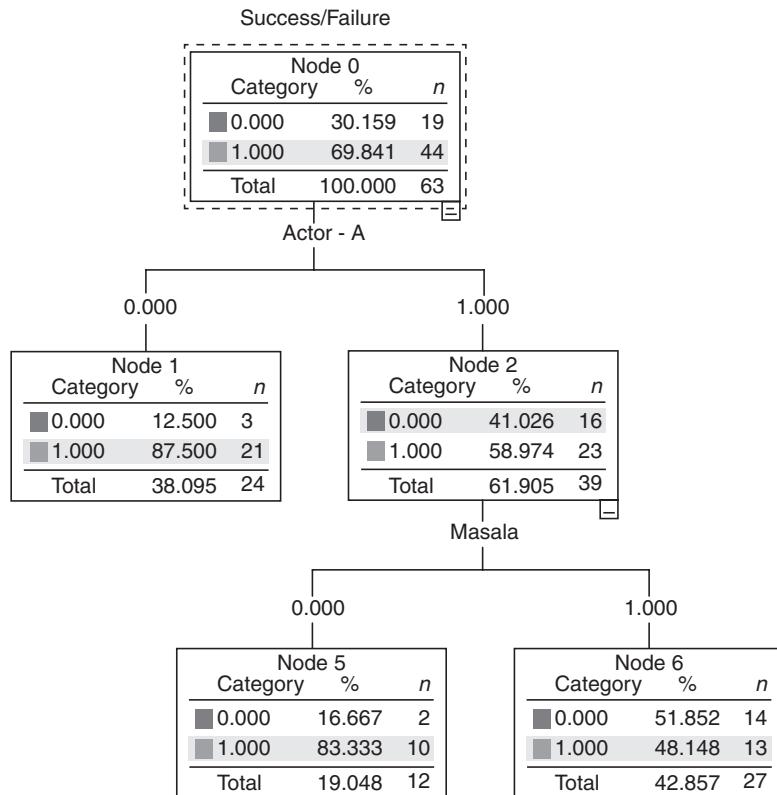
## EXERCISES

1. Use CART trees shown in Figures 12.10 and 12.11 to answer Questions (a) and (b). Figure 12.10 is a tree that has been developed to predict the success of a movie ( $Y = 1$ ) using the predictors budget and YouTube likes (YouTube-L). In Figure 12.11 the predators are actor category A (Actor\_A) and Movie Content Masala.



**FIGURE 12.10** Classification and regression tree with budget and YouTube likes as predictors.

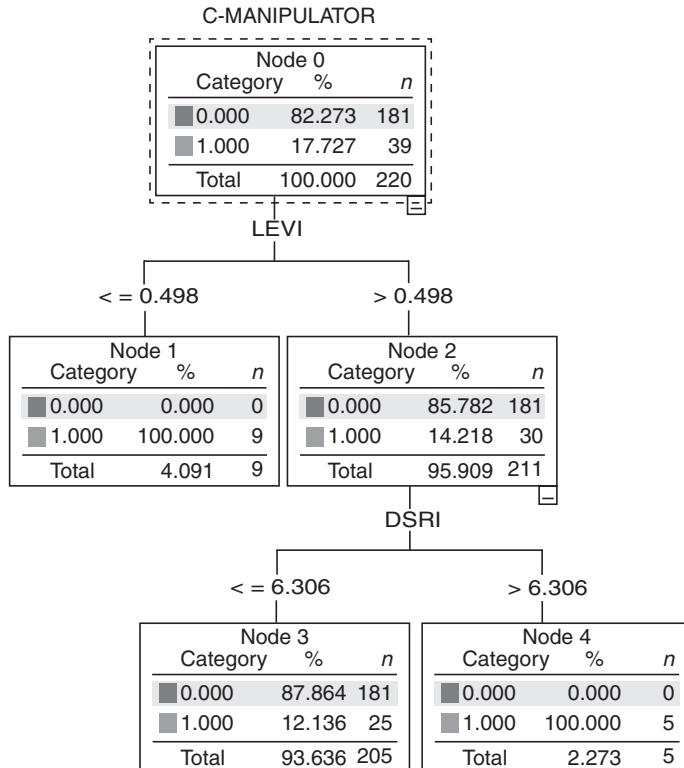
- (a) Calculate the Gini coefficient for node 0 in Figure 12.10. What is the change in the Gini coefficient between node 0 and nodes 1 and 2?
- (b) Create business rules based on the CART trees provided in Figures 12.10 and 12.11.
2. A CART tree is developed to identify earnings manipulators ( $Y = 1$ ) among Indian firms using financial indices:
- Days sales to receivables index (DSRI)
  - leverage index (LEVI)
- The tree is shown in Figure 12.12.



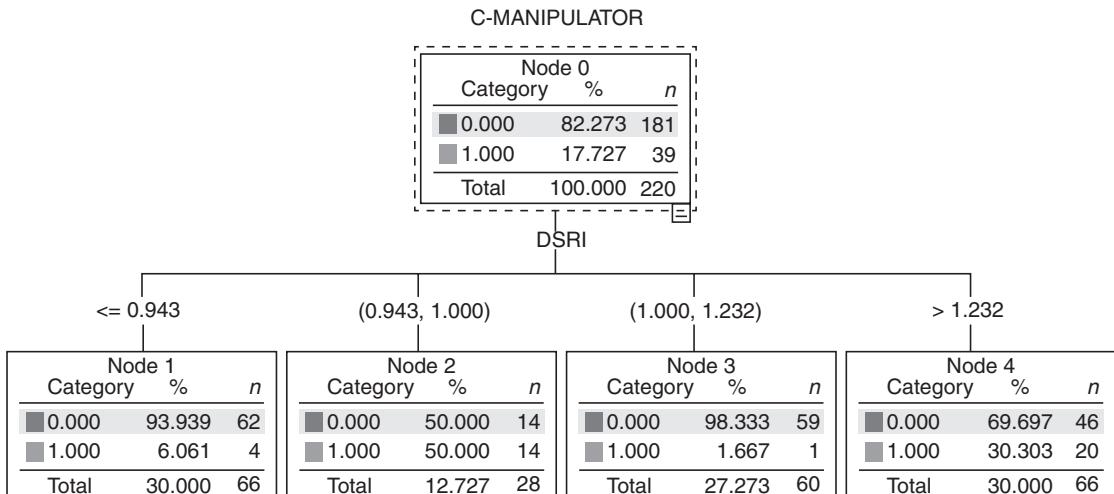
**FIGURE 12.11** Classification and regression tree with actor and genre as predictors.

- Calculate the total reduction in Gini index in the model (that is, between root node and all the leaf nodes).
  - Calculate the value of entropy at node 0.
  - Calculate the sensitivity and specificity of the CART model shown in Figure 12.12.
  - Write all the business rules that can be used for predicting earnings manipulator using the CART tree in Figure 12.12.
- The CHAID tree for the prediction of earnings manipulator ( $Y = 1$ ) is shown in Figure 12.13. Calculate the value of chi-square statistic and the corresponding  $p$ -value for the split used in the tree.
  - CHAID tree is developed to classify a tumour as benign ( $Y = 0$ ) or malignant ( $Y = 1$ ).<sup>3</sup> CHAID tree in Figure 12.14 is developed using the variable, perimeter-W (perimeter of the tumour).
    - Calculate the  $p$ -value of chi-square test of independence used to create the internal nodes.
    - What can you infer from the CHAID tree in Figure 12.14?
  - A CART tree for the prediction of cancer is developed using variables:
    - Radius-W,
    - Concave points-W,
    - Texture,
    - Texture-W, and
    - Concave points and is shown in Figure 12.15. Calculate the sensitivity and specificity of the CART Model Compare and Comment on nodes 6 and 10.

<sup>3</sup> Data source: <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29>



**FIGURE 12.12** CART tree for earnings manipulation.



**FIGURE 12.13** CHAID tree for earnings manipulations.

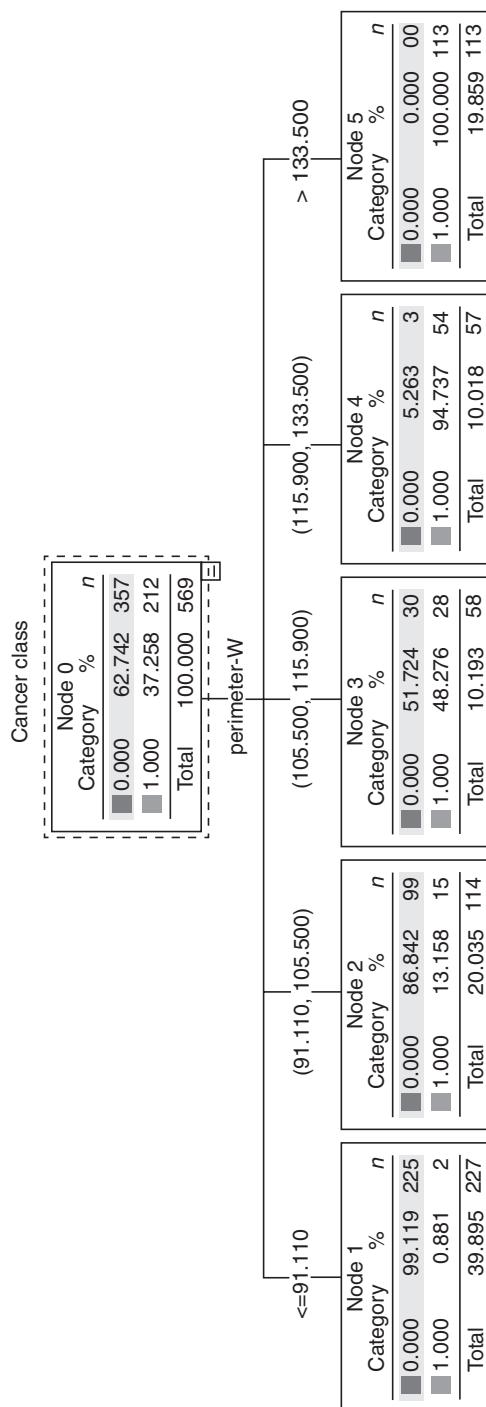


FIGURE 12.14 CHAID tree for cancer prediction.

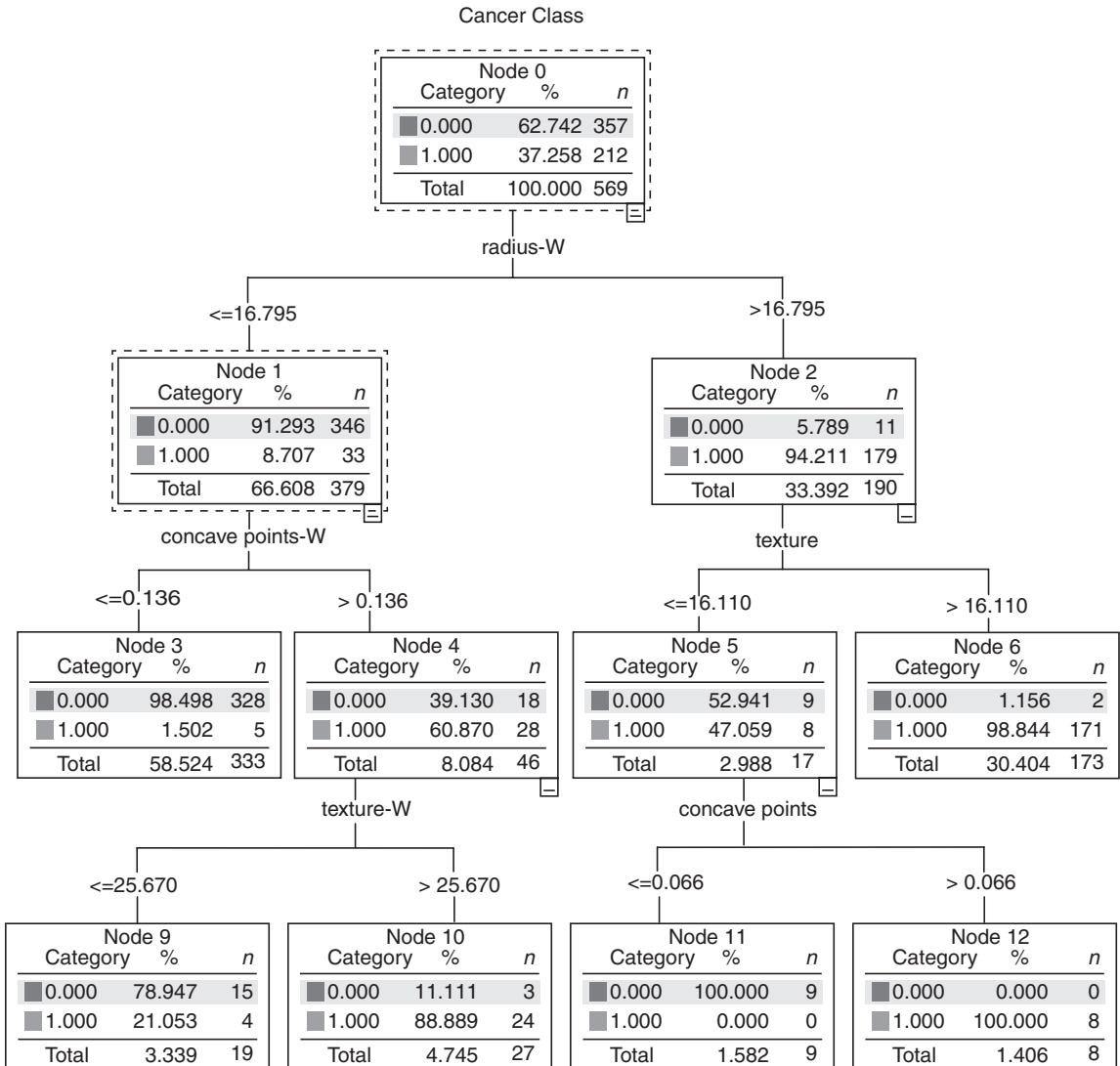


FIGURE 12.15 CART tree for cancer prediction.

## Breaking Barriers: Micro-Mortgage Analytics<sup>4</sup>

It had been a long day at the small shop that Naresh bhai (name changed to protect identity) owned at the street corner. As the day drew to a close, his thoughts wandered back to the previous day's developments. Finally, Shubham Housing Development Finance Company (hereinafter to be called 'Shubham'), a housing finance company, was willing to consider his application for a home loan. Naresh bhai felt hopeful yet apprehensive. He wondered if his dream of building a house for his family would finally come true. His reservations were not unfounded. As a teenager, he had moved to the city of Ahmedabad from a neighbouring village to seek better employment opportunities. Over time, he had managed to set up a petty shop where he sold daily provisions, condiments, and general merchandise. At the end of a good month, he would earn between INR 20,000 and 25,000 – a good enough household income for his family of four to have been able to afford a house. However, until the previous day, he had found it impossible to obtain a home loan from any of India's other housing finance companies. As a micro-entrepreneur, he fell outside the 'conventional' segment of loan applicants at these financial companies. He rarely, if ever, used to maintain any accounts of his business, and hence, he had been unable to produce the documents required by these housing finance companies such as tax returns, income proof, or bank statements. However, Shubham had seemed different and he felt optimistic about building his dream home after all.

Ajay Oak, the chief operating officer (COO) of Shubham, knew that in India there were 20 million home loan aspirants similar to Naresh bhai, who worked in the informal sector and constituted about 80% of India's urban residents.<sup>5</sup> Even though financing affordable housing was one of India's biggest challenges, it also offered a remarkable opportunity for enterprising finance companies. For Ajay and his team, a key challenge in serving this customer group was to effectively evaluate the loan repayment ability of prospective clients in the absence of basic mortgage documentation. Ajay's strong belief in the creditworthiness of the low-income group made him realize that the key to business success was to quickly identify potential customers through cost-effective means. Moreover, with the proliferation of new players in this nascent and rapidly expanding sector, providing faster application assessment time and differential processing fees for creditworthy clients would be vital for gaining a competitive edge and securing greater market share.

### Financing the Housing Aspirations of India's Informal Sector

In India, several million low-income families reside in crowded localities and endure a life of hardship in poorly constructed houses that have pitiful sanitary conditions. Hernando de Soto, the influential Peruvian economist, regards the provision of housing facilities to such families as the fundamental avenue to economic success and an improved quality of life. Indeed, many of these families aspire to live in good quality and affordable houses. However, securing a housing loan to realize their dreams has often proved to be exceedingly challenging.

<sup>4</sup> ©The Indian Institute of Management Bangalore. The case was written by Jitendra Rudravaram, Naveen Bhansali, Swetha Murthy, Sutirtha Roy, and U Dinesh Kumar and was distributed through Harvard Business Publishing. This case is not intended as an endorsement, source of primary data or to show effective or inefficient handling of decision or business processes. Reproduced with permission from IIM Bangalore.

<sup>5</sup> Source: Ministry of Housing and Urban Poverty Alleviation, Government of India.

## Case Study Continued...

With the traditional real estate sector experiencing a slowdown in India after 2009, real estate players rapidly invested in low-income housing projects ranging from INR 3 lakh up to INR 10 lakh (USD 1 = INR 62, as of November 2013). However, many of these low-income families were unable to access affordable mortgages from banking institutions and resorted to borrowing from moneylenders whose interest rates were anywhere between 36% and 60% per annum.<sup>6</sup>

Traditionally, banks and housing finance companies have viewed such applicants as high-risk profiles primarily because a vast majority of them lacked the basic documentation that would be necessary to even begin processing a loan application. Furthermore, these banks conventionally serve only high- or middle-income customers and have accordingly built their organizational capacity to cater to these market segments. Financing the low-income segment, however, demands a complete shift from an archetypal document-based underwriting process to an interview-based, on-field verification process. Such a process would put the financial institution right into the homes and workplaces of these customers, wherein the financial viability of a potential customer would be assessed through personal interviews, evaluation of the customer's workplace, and a thorough field-based due diligence.

Conventional housing finance institutions, which were more adept at traditional methods of assessment based on income and identification documents, were dismally short of the expertise needed to competently assess these customers. These challenges continued to keep traditional housing finance companies apprehensive about lending to low-income applicants and underscored their perception of low-income customers as a high-risk segment. At the same time, they recognized that the inadequacies in their loan underwriting process not only increased social inequity but also resulted in lost opportunities.

In 2010, the market for mortgages in the low-income housing bracket comprised more than 20 million households and was worth USD 182 billion.<sup>7</sup> Driven by this existing socio-financial opportunity, several new housing finance companies pioneered a 'small ticket' loan product to address the market gap. These companies realized that most of the applicants from this category would be ineligible for a housing loan if the conventional approach of document-based evaluation were applied. To address this challenge, companies such as Shubham introduced a new loan origination process that relied on detailed, field-based verification instead of the formal financial documentation process.<sup>7</sup>

### **Shubham Housing Development Finance Company: The Genesis and Rise**

Shubham Housing Development Finance Company commenced its lending operations in May 2011 through a single branch in New Delhi, with a vision to provide mortgage products and housing improvement loans to families that were excluded by traditional housing finance institutions. Led by Sanjay Chaturvedi as the CEO and Ajay Oak as the COO, Shubham became a pioneer in offering formal housing credit to low-income families from the informal sector and within 2 years had over

<sup>6</sup> Source: Pawan Gulani, Head of Sales and Product, Shubham Housing Development Finance Company.

<sup>7</sup> Source: Lalwani L, Merchant K and Venkatachalam B (2010), *Micro Mortgages: A Macro Opportunity in Low-Income Housing Finance*. Monitor Inclusive Markets, 2010. The study was commissioned by the NHB, funded by the FIRST initiative, and supported by the World Bank. Report available at <http://www.merchantkushagra.info/cloud/whitepaper-MicroMortgage-final-screen-10-19-10.pdf>

**Continued...**

40 branches spread across several cities in India. In November 2012, Gurgaon-based Shubham raised an additional USD 7.8 million (approximately INR 50 crore) from venture capitalists, two years after the housing finance firm had first raised around USD 2 million (approximately INR 12.5 crore).<sup>8</sup> By September 2013, Shubham had disbursed loans amounting to over INR 125 crore to around 2,300 applicants. Shubham's operating model sought to transcend the document-based underwriting process and instead followed an interview-based approach in order to understand an applicant's income and expense flows. A visual flow diagram of Shubham's loan approval process is provided in **Exhibit 1**.

### Loan Evaluation Process and Opportunities for Competitive Advantage

The interview-based approach allowed Shubham to assess the applicants based on their daily or monthly cash earnings as observed by Shubham's staff at the applicant's workplace, instead of relying on formal documentation for proof of their income. This relaxation of the documentation norms enabled deserving applicants to obtain a loan for purchasing/building a house. Most of these customers were typically employed as petty shop owners, taxi drivers, or household help.

The interview-based field assessment was conducted by a credit officer from Shubham who personally visited the applicant's residence and workplace in order to assess and verify the details provided by the applicant during the loan origination phase. The credit officer then created a story about the applicant's life by asking questions about his/her family, education, living conditions, income, expenses, liabilities, assets, work, and so on. In the instance of self-employed individuals, the credit officer would also spend a day with the applicant and observe the income and expenses incurred so as to create a Profit/Loss statement. Finally, the completed verification and assessment statement would result in a qualitative record of the person's ability to regularly service the loan. As expected, despite the benefits associated with field-based assessment, this method did present its own set of challenges for Shubham. The process of interviewing all customers irrespective of whether it resulted in a loan sanction increased the costs and the time expended on each customer. As per a 2010 study by Monitor on the Indian micro mortgage industry, the cost of originating a loan was sometimes be as high as 31% of the total transaction cost incurred.<sup>8</sup>

Moreover, the interview-based approach is a subjective evaluation of the customer's ability to service a loan. Therefore, the decision to approve or reject a loan is dependent on Shubham's field-based credit officers who would assess a customer's loan repayment ability. This business assessment is thus hinged on the credit officer's skill and subjective decision-making abilities. Comparing such decisions across different branches and field staff members in order to ensure consistency across branches also presents a significant challenge to Shubham's scale-up strategy. Often, the process of interviewing the candidates, verifying their credibility from third-party sources, and preparing the interview report leads to an extended loan processing time. Reducing this lead-time would allow Shubham to communicate its decisions to prospective customers quicker, thus serving them much better.

Shubham's customers range from contractual employees with the government to grocery vendors, from auto rickshaw drivers to semi-skilled laborers. These applicants are in the age group of

<sup>8</sup> Source: Bruhadeeswaran, R. 2012. *Shubham Housing Development Finance raises \$7.8M from Elevar, Helion, Others*. [www.vccircle.com/news/2012/11/12/shubham-housing-development-finance-raises-78m-elevar-helion-others](http://www.vccircle.com/news/2012/11/12/shubham-housing-development-finance-raises-78m-elevar-helion-others)

## Case Study Continued...

22 to 63 years. They hail from different parts of the country and include school dropouts as well as post-graduates. Thus, Shubham's informal sector customers belong to a wide socioeconomic and demographic pool. A better understanding of these customer groups would enable Shubham to target its marketing efforts towards those customer segments that would lead to faster loan processing.

Shubham's data acquisition and collection processes at every stage of applicant assessment are among its greatest sources of competitive advantage. Ajay envisioned that an application-scoring model built using the data generated from the field-level interactions would enhance decision making at the branch level.

The aforementioned Monitor study on micro mortgage sector reports that as per industry estimates, for a loan of INR 400,000, loan origination would cost INR 8,000. This constitutes 31% of the total transaction cost and 2.8% of the total loan disbursed. In addition to the loan origination cost, a non-performing asset (NPA) provision of INR 4,000 and a loan servicing cost of INR 13,844 brought the total transaction cost<sup>7</sup> to INR 25,844 (**Exhibit 2**). Upon sanction of the loan, the loan applicant would be charged a non-refundable amount equivalent to 2.5% of the sanctioned amount towards loan processing fees. In this instance, it was INR 10,000.

Evaluation of the loan sanction probability early in the process by the credit team could reduce field costs – the 31% cost component incurred in every transaction – by identifying creditworthy customers before the credit officer's visit (**Exhibit 3**). The product team could also make the product more competitive by offering differential loan processing charges to high potential clients (**Exhibit 4**).

Additionally, the sales team could reduce the operation costs that result from pursuing all prospective leads and could source better clients from demand-side aggregators as the business grew. The model would also reduce incorrect sanctions of more risky loan applications owing to subjective evaluations made by credit officers and would thereby help the standardization of decision making and reporting across nationwide branches.

## Data Analysis

For the application scoring model, Ajay's analytics team acquired field-level data consisting of variables collected from each applicant (**Exhibit 5**). In addition to the socioeconomic variables (such as age, marital status, education, housing situation, income, expenses, loan amount requested, and so on) that were gathered from the application forms, Ajay's team calculated the standard mortgage ratios for each applicant. These ratios were **Installment to Income Ratio (IIR)**, **Installment to Disposable Income Ratio (IAR)**, **Fixed Obligation to Income Ratio (FOIR)**, **Loan to Value Ratio (LTV)**, and **Loan to Cost Value Ratio (LVR)** (**Exhibit 6**).

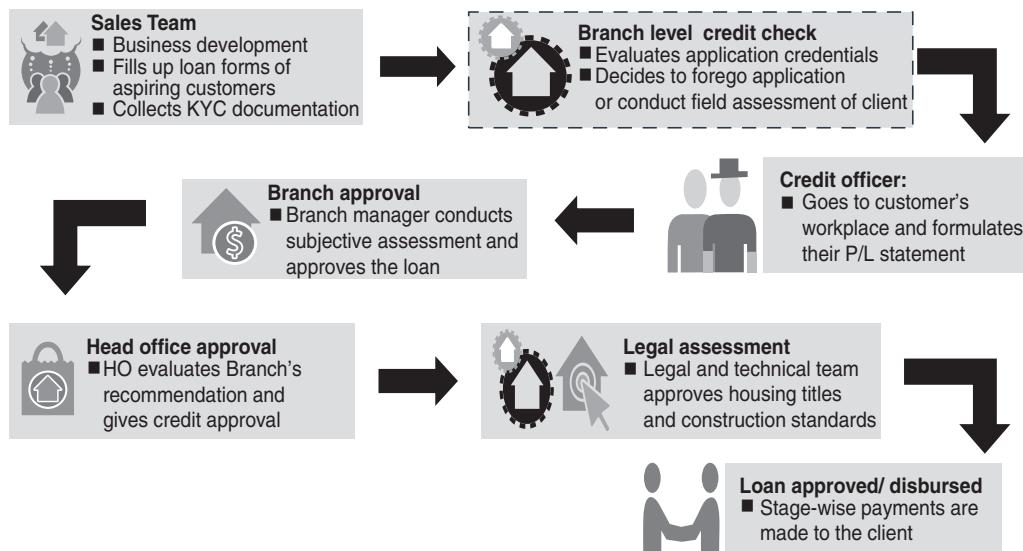
These ratios helped to capture important factors such as the ability of an applicant to repay debt, the proportion of property value sought as loan, and so on. Since these ratios are used extensively in processing applications, these were included in the data set while building the scoring models so as to ascertain whether these ratios had a higher significance than the constituent individual variables.

The techniques of Chi-squared Automatic Interaction Detection (CHAID) and binomial logistic regression were employed to predict the loan sanctions. **Exhibit 7** contains the Tier classification of Indian cities. The CHAID model is depicted in **Exhibit 8** and the classification results obtained

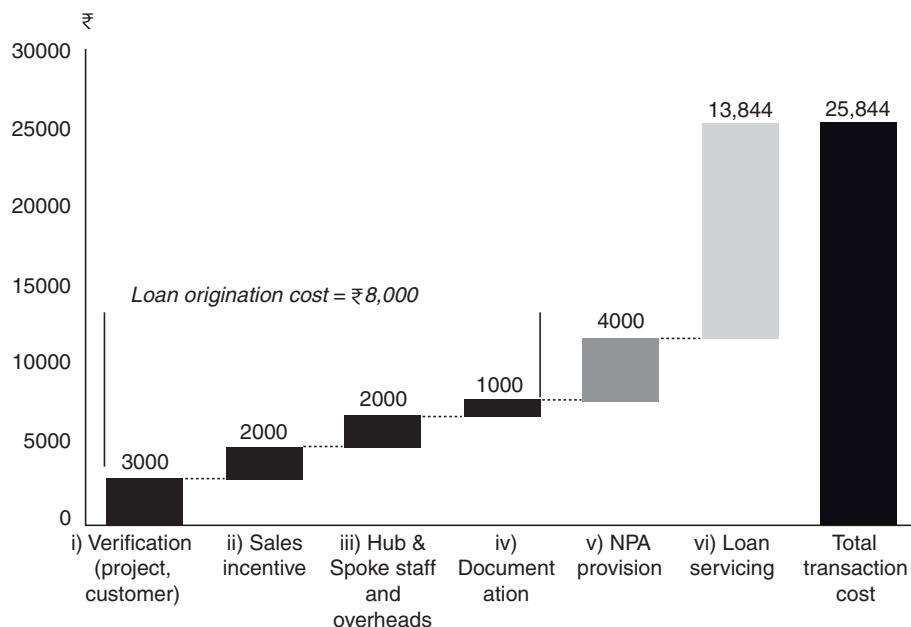
Case Study

**Continued...**

through the model for the training and the validation data sets are shown in **Exhibits 9** and **10**, respectively. **Exhibit 11** contains a sample of 100 applications used for testing the CHAID model.



**EXHIBIT 1** Shubham's loan approval process.



**EXHIBIT 2** Cost split-up for an INR 4 lakh loan at ROI 14% and 10% cost of capital. Source: monitor inclusive markets (2010).

Continued...

**EXHIBIT 3** Impact of cost reduction

Cost Components	Current	Model Application
Loan requested	4,00,000	4,00,000
Loan origination cost	8,000	4,500
Total transaction cost	25,844	22,844
% of transaction cost	6.46%	5.59%
% Origination cost vs. Transaction cost	30.95%	20.14%

Source: Monitor inclusive markets (2010).

*Current situation:* For a rejected customer loan request of **4 lakh @ 14% ROI**

*Loan origination:* INR **8,000** (3,000 for client and project verification + 2,000 for sales incentives + 2,000 for staff and overhead expenses + 1,000 for documentation) Other costs = 0

**Total cost = 8,000**

*On applying analytical model:* For a rejected customer loan request of **4 lakh @ 14% ROI**

*Loan origination:* INR **4,500** (1,000 for client and project verification + 1,000 for sales incentives + 2,000 for staff and overhead expenses + 500 for documentation) Other costs = 0

**Total cost = INR 4,500**

**EXHIBIT 4** Illustrative example of differential processing fee

**Identical Fee Structure**

Customer	Customer 1	Customer 2	Customer 3
Loan amount required	400,000	400,000	400,000
Processing rate charged	2.50%	2.50%	2.50%
Processing fee charged	10000	10000	10000
Total processing fee	30000		

**Differential Fee Structure**

Customer	Customer 1	Customer 2	Customer 3
Applicant score	0.95	0.85	0.75
Loan requested	400,000	400,000	400,000
Processing rate charged	2%	2.50%	3%
Processing fee charged	8000	10000	12000
Total processing fee	30000		

*Note:* Classification of sanctioned customers was highest when their probabilities of sanction were greater than 0.9 (Probability of sanctions > 0.9 customers => More sanctions and less rejections). From a strategic point of view, these customers could be offered a lower processing fee as compared to the rest of the applicants.

**Continued...**

**EXHIBIT 5** Data variables

Variable	Definition
ID	Unique Identifier for each application
Decision	Credit decision taken for the applicant 1 = Sanction, 0 = Reject
Build_Selfcon	Variable to indicate whether applicant seeks a home loan for self-construction or a builder-promoted project
Selfcon_code	If Build_Selfcon = 'Self Construction', then Selfcon_code = 1; if Build_Selfcon = 'Builder', then Selfcon_code = 0
Tier	City tier where the loan was sought. Tier-1 = Major City, Tier-2 = Minor City, Tier-3 = Town/Village
Tier_1	If Tier = 'Tier-1', Tier_1 = 1; else Tier_1 = 0
Tier_2	If Tier = 'Tier-2', Tier_2 = 1; else Tier_2 = 0
Accommodation_Class	Variable to indicate whether applicant resides currently in rented or non-rented premises
Accoclass	If Accommodation_Class = 'Rented', Accoclass = 1, else Accoclass = 0
Loan_Type	Variable to indicate if loan was sought for Home loan or Home Improvement loan
Loantype	If Loan_Type = 'Home _Loan', Loantype = 1; else Loantype = 0
Gender	Applicant's Gender
Sex	If Gender = 'Male', Sex = 1; else Sex = 0
Employment_Type	Variable to indicate whether the applicant was salaried or self-employed
Etype	If Employment_Type = 'Self_Employed', Etype = 1; else Etype = 0
doc_proof_inc	Indicates whether the applicant submitted documentary proof of income
Docprf	If doc_proof_inc = 'Y', Docprf = 1; else Docprf = 0
Marital_Status	Indicates if applicant is married or single currently
Marstat	If Marital_Status = 'Married', Marstat = 1; else Marstat = 0
Employer_Type	Applicant's Employer's category (Business, Corporate, Government, Ind/Small Business)
Emp_type_1	If Employer_Type = 'Business', Emp_type_1 = 1; else Emp_type_1 = 0
Emp_type_2	If Employer_Type = 'Govt', Emp_type_2 = 1; else Emp_type_2 = 0
Emp_type_3	If Employer_Type = 'Corporate', Emp_type_3 = 1; else Emp_type_3 = 0
Education_Class	Education of the applicant
Educlass_2	If Education_Class = 'GRADUATE+', Educlass_2 = 1; else Educlass_2 = 0
Educlass_1	If Education_Class = 'UNDERGRADUATE', Educlass_1 = 1; else Educlass_1 = 0
Mode_of_origin_class	The source from which the application originated
Oriclass_1	If mode_of_origin_class = 'Reference', Oriclass_1 = 1; else Oriclass_1 = 0
Oriclass_2	If mode_of_origin_class = 'Own database field visit', Oriclass_2 = 1; else Oriclass_2 = 0
eom_25	Variable to indicate whether the application was received after the 25 <sup>th</sup> of the month

Continued...

**EXHIBIT 5** Data variables—Continued

Variable	Definition
oldemi_d	Variable to indicate if applicant had old loans
bs_d	Variable to indicate if applicant has bank savings
Age	Age of applicant
Yrsadd	Years in current residential address
Yrsjob	Years in current job
Expen	Monthly expenses of applicant
Totinc	Monthly income of applicant
Dispinc	Total monthly income – Total monthly expenses
Marval	Market value of the property for which loan is sought
Oldemi	EMI for earlier loans that the applicant pays every month
Loanreq	Loan amount requested by applicant
Term	Term for the loan
Dwnpay	Down payment by applicant
Banksave	Bank saving of applicants
Calcemi	EMI calculated for the applicant's requested loan amount
IIR	Calcemi/Monthly total income
Gender	Applicant's Gender
IAR	Calcemi/Monthly disposable income
FOIR	(Calcemi + Oldemi)/Total monthly income
LTV	Total loan requested/Market Value
LVR	Total loan requested/Property registered value
dwnp_prop	Dwnpay / (Dwnpay + Loanreq)
mfoir_p	((Oldemi + Calcemi) * 100) Dispinc
dwnp_prop_p	Dwnp_prop * 100
dispinc_s	Dispinc/10000
marval_s	Marval/1000000
loanreq_s	Loanreq/100000
banksave_s	Banksave/10000
calcemi_s	Calcemi/10000
oldemi_s	Oldemi/10000
Tier_2XAccoclass	Interaction variable of Tier_2 & Accoclass

Continued...

**EXHIBIT 6** Standard mortgage ratios

IIR	$\frac{\text{Equated Monthly Installments (EMI)}}{\text{Total Household Income}}$
IAR	$\frac{\text{Equated Monthly Installments (EMI)}}{\text{Disposable Income}}$ Disposable Income = Total Household Income – Total Expenses
LTV	$\frac{\text{Total Loan Requested}}{\text{Market Value of Property}}$
LVR	$\frac{\text{Total Loan Requested}}{\text{Property Value}}$ Property value is registered value of property at the municipality
FOIR	$\frac{\text{EMI} + \text{Ongoing Loan EMI}}{\text{Total Household Income}}$

**EXHIBIT 7** Tier classification of Indian cities

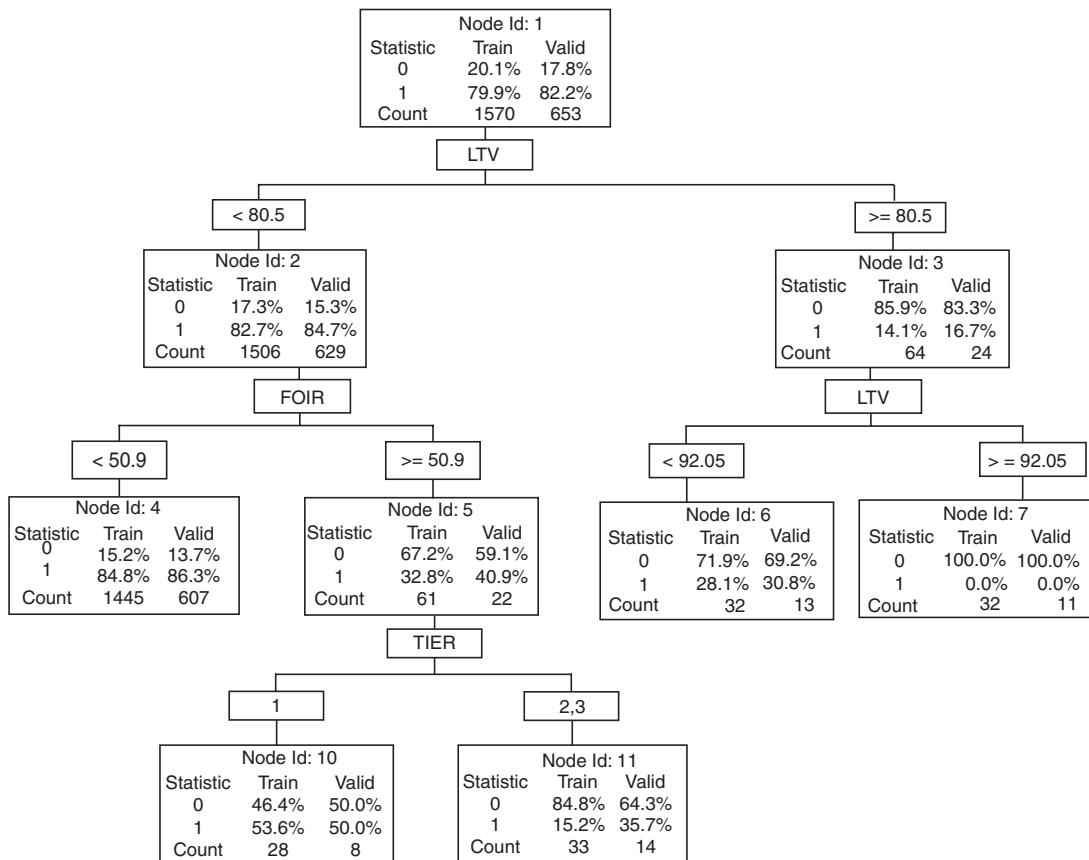
Tier Classification	City	Tier Classification	City	Tier Classification	City
1	Bangalore	2	Vadodara	2	Aurangabad
1	Chennai	2	Ludhiana	2	Srinagar
1	Delhi	2	Agra	2	Bhilai
1	Hyderabad	2	Meerut	2	Rajahmundry
1	Kolkata	2	Nashik	2	Kakinada
1	Mumbai	2	Faridabad	2	Nellore
1	Ahmedabad	2	Varanasi	2	Solapur
2	Bhopal	2	Jabalpur	2	Ranchi
2	Kanpur	2	Jamshedpur	2	Guwahati
2	Jaipur	2	Allahabad	2	Gwalior
2	Nagpur	2	Amritsar	2	Chandigarh
2	Lucknow	2	Indore	2	Patiala
2	Patna	2	Gorakhpur	2	Jodhpur
2	Pune	2	Hubli-Dharwad	2	Tiruchirapalli
2	Visakhapatnam	2	Raipur	2	Salem

Continued...

**EXHIBIT 7** Tier classification of Indian cities—Continued

Tier Classification	City	Tier Classification	City	Tier Classification	City
2	Kochi	2	Mangalore	2	Rajkot
2	Madurai	2	Belgaum	2	Cuttack
2	Coimbatore	2	Guntur	2	Amravati
2	Warangal	2	Bhubaneswar	2	Pondicherry
2	Surat	2	Bhavnagar	3	All other cities
2	Asansol				

Source: Sixth central pay commission of India.



**EXHIBIT 8** Chi-squared automatic interaction detection (CHAID) model.

Continued...

**EXHIBIT 9** Classification results of CHAID model for training data set

Target	Outcome	Frequency Count
0	0	83
1	0	14
0	1	232
1	1	1241

Classification Table: Training

Sensitivity	98.8%
Specificity	26.3%

**EXHIBIT 10** Classification results of CHAID model for validation data set

Target	Outcome	Frequency Count
0	0	29
1	0	9
0	1	87
1	1	528

Classification Table: Training

Sensitivity	98.3%
Specificity	25%

**EXHIBIT 11** Sample data set used to test the CHAID model

Id	Decision	Tier of City	Age	Yrsadd	Oldemi	FOIR	LTV
DEL-S6QA-442656	Reject	1	37	13	0	32.00	57.00
GUJ-KE2E-383829	Reject	1	45	9	0	25.65	177.78
MDG-TR7G-885286	Reject	3	27	2	12679	56.54	67.00
MDG-EA6U-239956	Reject	3	32	5	0	31.00	89.00
GUJ-CHA6-377655	Reject	1	28	6	0	51.00	71.00
DEL-CRU4-638996	Reject	1	39	8	0	30.78	66.67
RAJ-CU3A-435927	Reject	2	37	5	2240	49.64	73.00
SHB-PRU8-975396	Reject	3	37	6	0	49.00	80.00
NAG-HE5R-459425	Reject	2	34	6	0	32.00	25.00
GUJ-9ASP-257984	Reject	1	31	25	0	54.00	80.00
MDG-SE8U-743796	Reject	3	45	2	0	47.00	59.00
FBD-GUH2-767994	Reject	2	35	10	0	34.00	72.00

Continued...

**EXHIBIT 11** Sample data set used to test the CHAID model—Continued

<b>Id</b>	<b>Decision</b>	<b>Tier of City</b>	<b>Age</b>	<b>Yrsadd</b>	<b>Oldemi</b>	<b>FOIR</b>	<b>LTV</b>
DEL-5UBR-632944	Reject	1	29	10	0	52.07	145.45
SAH-P6AW-477253	Reject	3	50	1	0	34.00	62.00
GUJ-9RAS-857253	Reject	1	40	4	0	35.63	120.94
JAM-4RES-776444	Reject	3	30	28	40000	30.99	50.00
GUJ-ME7A-946978	Reject	1	24	23	0	38.00	85.00
VAD-QA7H-459898	Reject	2	27	10	0	27.00	40.00
RAJ-CHAS-963229	Approve	2	43	10	0	49.00	78.00
SRT-CR8J-295523	Approve	2	33	6	0	31.00	51.00
GUJ-9HUS-878464	Approve	1	28	2	0	47.00	49.00
SHB-9UDA-632993	Approve	3	33	10	0	38.00	26.00
MDG-WE7R-964695	Approve	3	41	1	0	22.00	40.00
SHB-VUH7-746557	Approve	3	47	14	0	37.00	62.00
SRT-6RES-429394	Approve	2	32	1	0	42.00	67.00
SRT-WE2U-246244	Approve	2	39	3	0	37.00	61.00
GUJ-PR8Z-554775	Approve	1	52	4	0	48.00	35.00
MRT-SP2P-623473	Approve	2	41	10	0	14.00	29.00
GUJ-CRA8-753769	Approve	1	36	3	2455	43.86	46.00
SHB-Q2FA-784777	Approve	3	36	10	0	47.00	80.00
GUJ-2RAH-276674	Approve	1	34	31	0	35.00	23.00
PUN-9ADA-922444	Approve	2	27	3	0	50.00	65.00
SRT-7AWR-894639	Approve	2	30	3	0	39.00	53.00
IDR-K5BR-529575	Approve	2	27	10	0	44.00	24.00
SHB-TR5G-853778	Approve	3	32	20	5495	43.39	80.00
AMR-PET7-268923	Approve	2	27	26	8743	47.10	48.00
SHB-J8TE-869465	Approve	3	33	3	0	35.00	62.00
NAG-SUM5-824386	Approve	2	32	30	0	41.00	77.00
SRT-MEP2-939449	Approve	2	32	4	0	47.00	63.00
GUJ-B3SP-584592	Approve	1	39	7	3310	47.67	80.00
AJM-DR3F-728347	Approve	3	49	3	0	23.00	45.00
SRT-GEX5-278492	Approve	2	41	3	0	42.00	32.00
MDG-TR8Y-794363	Approve	3	28	7	4772	45.79	39.00
SRT-V7MU-562553	Approve	2	26	1	0	38.00	66.00

(Continued)

**Continued...****EXHIBIT 11** Sample data set used to test the CHAID model—Continued

<b>Id</b>	<b>Decision</b>	<b>Tier of City</b>	<b>Age</b>	<b>Yrsadd</b>	<b>Oldemi</b>	<b>FOIR</b>	<b>LTV</b>
SRT-FRE8-682823	Approve	2	35	4	0	47.00	55.00
SHB-Y5DR-546357	Approve	3	35	1	0	45.00	27.00
SRT-PR5B-887235	Approve	2	31	2	0	14.00	46.00
MDG-P8UH-689386	Approve	3	39	6	6741	49.06	61.00
MRT-HU8A-244932	Approve	2	52	1	0	47.00	36.00
SHB-AI9Q-623595	Approve	3	29	8	0	25.00	79.00
GUJ-SWA7-532964	Approve	1	48	4	0	31.00	12.00
FBD-S2UC-952959	Approve	2	30	3	0	30.00	42.00
PUN-2HAS-626445	Approve	2	29	2	2116	50.05	67.00
RAJ-TR4T-834863	Approve	2	24	2	0	43.00	63.00
DEL-TRA3-376247	Approve	1	40	10	0	23.00	62.00
DEL-BE3P-484634	Approve	1	33	11	0	43.00	50.00
GUJ-DUC2-753522	Approve	1	35	2	0	40.00	53.00
NAS-TEQ8-493685	Approve	2	25	5	0	45.00	36.00
GUJ-8UNA-562324	Approve	1	47	0	0	33.00	73.00
GUJ-8RAY-382542	Approve	1	35	5	0	7.00	30.00
SRT-P9UF-599895	Approve	2	33	4	0	36.00	53.00
SHB-QEC9-262767	Approve	3	31	2	20220	44.60	80.00
DEL-F9SP-826567	Approve	1	40	20	0	19.00	27.00
MDG-Y8PH-666948	Approve	3	33	30	0	44.00	55.00
SHB-WR8C-828893	Approve	3	33	5	0	21.00	62.00
NAS-H7BR-796623	Approve	2	30	10	0	46.00	33.00
MDG-VA5R-522867	Approve	3	41	1	0	41.00	28.00
GUJ-PHE2-945563	Approve	1	33	9	0	43.00	55.00
SHB-3ATH-365745	Approve	3	27	2	0	50.00	50.00
RAJ-S5UP-432944	Approve	2	33	8	0	49.00	58.00
SHB-GAJ5-656633	Approve	3	27	3	17349	55.41	33.00
DEL-6HAP-992534	Approve	1	39	2	0	9.00	71.00
SRT-XER8-873652	Approve	2	34	1	7141	49.12	22.00
IDR-7HUX-947934	Approve	2	43	4	0	34.00	32.00
MRT-Y5TH-366839	Approve	2	31	2	0	39.00	71.00

Continued...

**EXHIBIT 11** Sample data set used to test the CHAID model—Continued

Id	Decision	Tier of City	Age	Yrsadd	Oldemi	FOIR	LTV
SHB-7UWA-942266	Approve	3	48	3	0	40.00	37.00
SRT-FEW6-377924	Approve	2	42	2	0	46.00	56.00
SRT-HU4H-759456	Approve	2	28	8	0	44.00	24.00
IDR-PR2B-373527	Approve	2	54	20	0	27.00	89.00
RAJ-NE5W-936659	Approve	2	51	49	0	49.00	42.00
GUJ-G2BR-585643	Approve	1	25	20	0	31.00	46.00
VAD-9EGU-733365	Approve	2	39	1	0	31.00	43.00
DEL-TR5G-982679	Approve	1	57	24	0	39.00	62.00
SRT-CEV2-858888	Approve	2	35	1	11369	49.55	52.00
SHB-7EFR-562262	Approve	3	45	1	0	26.00	60.00
VAD-5REP-436572	Approve	2	34	30	1493	44.73	79.00
GUJ-TR2X-454675	Approve	1	37	1	2172	46.89	69.00
MRT-FR4H-849474	Approve	2	34	5	0	39.00	57.00
SRT-VAG4-472429	Approve	2	25	8	0	46.00	47.00
DEL-SP7T-869623	Approve	1	60	33	0	21.00	14.00
SHB-9ABR-357338	Approve	3	31	2	0	40.00	50.00
DEL-6HED-235585	Approve	1	28	1	0	38.00	75.00
PUN-STA6-298789	Approve	2	37	25	1558	44.41	42.00
DEL-S3ET-655923	Approve	1	32	5	0	28.00	80.00
AJM-BRU8-457322	Approve	3	38	8	0	49.00	34.00
DEL-REC6-289437	Approve	1	42	2	0	44.00	67.00
RAJ-SPE8-423236	Approve	2	42	2	0	50.00	71.00
IDR-THE9-644793	Approve	2	40	15	0	33.00	44.00
JAM-NU9R-969974	Approve	3	27	24	0	43.00	47.00
GUJ-G6VU-427765	Approve	1	33	20	0	37.00	52.00

#### CASE QUESTIONS

- Examine the decision tree in **Exhibit 8** and come up with a strategy which Ajay and his team should adopt to identify creditworthy customers early in the assessment phase.

**Continued...**

2. Using the classification table in **Exhibit 10**, answer the below questions:(a) Given 100 loan applicants who have been approved a loan, how many of them would the model be able to correctly predict as sanctions? (b) Calculate the proportion of actual rejects which have been incorrectly classified as sanctions.
3. Apply the CHAID decision tree on the 100 data points provided in **Exhibit 11** and construct the classification table. Is it similar to the one provided in **Exhibit 10**? What are your conclusions regarding the CHAID decision tree results?
4. Examine the decision tree in **Exhibit 8** and explain why derived variables such as LTV and FOIR have better explanatory power than the individual variables used to derive them.
5. Develop a CART model to predict whether an applicant should be given loan and generate business rules.
6. Develop a Random Forest Model using the case data. Compare the model accuracy based on sensitivity and specificity of CART and CHAID with Random Forest.

**REFERENCES**

1. Armstrong R A (2014), “When to Use Bonferroni Correction”, *Ophthalmic and Physiological Optics*, 34, 502–508.
2. Breiman L, Friedman J H, and Olshen R A and Stone C J (1984), “Classification and Regression Trees”, Chapman and Hall, USA
3. Kass G V (1980), “An Exploratory Technique for Investigating Large Quantities of Categorical Data”, *Applied Statistics*, 20(2), 119–127.

# 13

# Forecasting Techniques

“Those who have knowledge don’t predict. Those Who Predict Don’t have Knowledge”.

— Lao Tzu

## LEARNING OBJECTIVES

- LO 13-1** Understand the importance of forecasting and its impact on the effectiveness of the supply chain and overall performance of an organization.
- LO 13-2** Learn various components of time-series data such as trend, seasonality, cyclical component, and random component.
- LO 13-3** Learn different techniques such as moving average, exponential smoothing, and Croston’s method.
- LO 13-4** Learn Auto-Regression (AR), Moving Average (MA), and Auto-Regressive Integrated Moving Average models (ARIMA).
- LO 13-5** Learn practical challenges associated with forecasting models using case studies.

## IMPORTANCE OF FORECASTING

Forecasting is one of the most important and frequently addressed problems in analytics. Inaccurate forecasting can have significant impact on both top line and bottom line of an organization. For example, non-availability of product in the market can result in customer dissatisfaction, whereas, too much inventory can erode the organization’s profit. Thus, it becomes necessary to forecast the demand for a product and service as accurately as possible.

## 13.1 | INTRODUCTION TO FORECASTING

Forecasting is by far the most important and frequently used application of predictive analytics since it has significant impact on both top line and bottom line of an organization. Every organization prepares long-range and short-range planning for the organization and forecasting demand for product and service is an important input for both long-range and short-range planning. Different capacity planning problems such as manpower planning, machine capacity, warehouse capacity, materials requirements planning (MRP) will depend on the forecasted demand for the product/service. Budget allocation for marketing promotions and advertisement are usually made based on forecasted demand for the product. Forecasting can be very challenging due to several factors that

influence the demand and scale of business with stock keeping units (SKUs) running into several millions. For example:

1. Boeing 747-400 has more than 6 million parts and several thousand unique parts (Hill, 2011). Forecasting demand for spare parts is important since non-availability of mission critical parts can result in aircraft on ground (AOG) which can be very expensive for airlines.
2. Amazon.com sells more than 350 million products through its E-commerce portal. Amazon itself sells about 13 million SKUs and has more (about 2 million) retailers selling their products through Amazon (Ali, 2017). Predicting demand for these products is important since overstocking can impact the bottom line and under stocking can result in customer dissatisfaction. Amazon.com may not stock all SKUs they sell through their portal since most of them are sold by their suppliers (online marketplace) directly to the customers, but even if they have to predict demand for products directly sold by them, then the number of SKUs is 13 million.
3. Walmart sells more than 142,000 products through their supercenters (*source: Walmart website<sup>1</sup>*). Being a brick-and-mortar retail store, Walmart does not have the advantages of Amazon.com (being also a market place, Amazon do not have to predict demand for all the products sold through their portal). They have to maintain stock for each and every product sold by Walmart and predict demand for the products as accurately as possible.
4. Demand for products and service is not the only application of forecasting, even manpower planning requires the use of sophisticated models. Indian information technology (IT) companies struggle to manage the right level of manpower for each skill required to manage their business. This would involve forecasting business opportunities, skills required to manage current and future projects, and so on.
5. Many products may have intermittent demands, that is, gap between two demands can be long and the gap itself may be random. The modeler has to forecast the next instance of demand and the actual demand quantity when demand occurs, making it much more difficult to forecast.
6. One of the innovative applications of forecasting was the Netflix forecasting contest in which the participants were challenged to forecast the movie rating (on a scale of 1 to 5) likely to be given by a customer for a movie. An accurate customer movie rating forecast can further be used for movie recommendations to customers.

## 13.2 | TIME-SERIES DATA AND COMPONENTS OF TIME-SERIES DATA

Time-series data is a data on a response variable,  $Y_t$ , such as demand for a spare parts of a capital equipment or a product or a service or market share of a brand observed at different time points  $t$ . Whenever we have a forecasting problem, we will be using a time-series data. The variable  $Y_t$  is a random variable. The data points or measurements are usually collected at regular intervals and are arranged in chronological order. If the time-series data contains observations of just a single variable (such as demand of a product at time  $t$ ), then it is termed as univariate time series. If the data consists of more than one variable, for example, demand for a product at time  $t$ , price at time  $t$ , amount of money spent by the company on promotion at time  $t$ , competitors price at time  $t$ , etc. then it is called multivariate time-series data.

<sup>1</sup> Source: [http://corporate.walmart.com/\\_news/\\_news-archive/2005/01/07/our-retail-divisions](http://corporate.walmart.com/_news/_news-archive/2005/01/07/our-retail-divisions)

From a forecasting perspective, a time-series data can be broken into the following components [Figures 13.1(a)–(d)]:

1. **Trend Component ( $T_t$ ):** Trend is the consistent long-term upward or downward movement of the data over a period of time.
2. **Seasonal Component ( $S_t$ ):** Seasonal component (measured using seasonality index) is the repetitive upward or downward movement (or fluctuations) from the trend that occurs within a calendar year such as seasons, quarters, months, days of the week, etc. The upward or downward fluctuation may be caused due to festivals, customs within a society, school holidays, business practices within the market such as '*end of season sale*', and so on. For example, in India demand for many products surge during the festival months of October and November. A similar pattern exists during December in many countries due to Christmas. Usually, for a given context seasonal fluctuation occurs at fixed intervals (such as months, quarters) known as periodicity of seasonal variation and repeats over time.
3. **Cyclical Component ( $C_t$ ):** Cyclical component is fluctuation around the trend line that happens due to macro-economic changes such as recession, unemployment, etc. Cyclical fluctuations have repetitive patterns with a time between repetitions of more than a year. Whereas in case of seasonality, the fluctuations are observed within a calendar year and are driven by factors such as festivals and customs that exist in a society. A major difference between seasonal fluctuation and cyclical fluctuation is that seasonal fluctuation occurs at fixed period within a calendar year, whereas cyclical fluctuations have random time between fluctuations. That is, periodicity of seasonal fluctuations is constant, whereas the periodicity of cyclical fluctuations is not constant.
4. **Irregular Component ( $I_t$ ):** Irregular component is the white noise or random uncorrelated changes that follow a normal distribution with mean value of 0 and constant variance.

The time-series data can be modelled as an addition of the above components or product of the above components. The additive time-series model is given by

$$Y_t = T_t + S_t + C_t + I_t \quad (13.1)$$

The additive models assume that the seasonal and cyclical components are independent of the trend component. Additive models are not very common since in many cases the seasonal component may not be independent of trend. At the Indian Institute of Management Bangalore (IIMB) there are many weekend programs and the number of students enrolled in these programs is fixed. The demand for food at the canteens of IIMB increases by a fixed quantity on Saturdays. This increase for demand is additive in nature.

The multiplicative time-series model is given by

$$Y_t = T_t \times S_t \times C_t \times I_t \quad (13.2)$$

Multiplicative models are more common and are a better fit for many data sets. In many cases, we will use the form  $Y_t = T_t \times S_t$  which is simpler form of Eq. (13.2). To estimate the cyclical component we will need a large data set.

The additive model is appropriate if the seasonal component remains constant about the level (or mean) and does not vary with the level of the series, while the multiplicative model is more appropriate if seasonal variation is correlated with level (local mean).

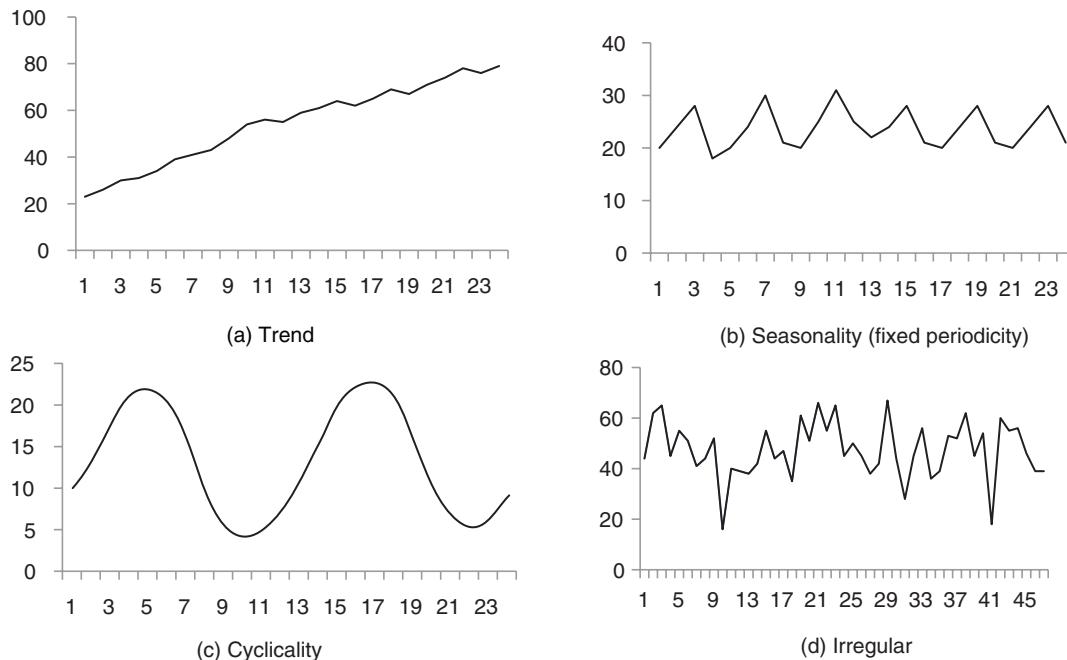


FIGURE 13.1 Trend in time-series data.

### 13.3 | FORECASTING TECHNIQUES AND FORECASTING ACCURACY

There are many forecasting techniques developed based on different logics. Simple techniques such as moving average and exponential smoothing predict the future value of a time-series data as a function of the past observations. Whereas the regression-based models such as auto-regressive (AR), moving average (MA), auto-regressive and moving average (ARMA), auto-regressive integrated moving average (ARIMA), and auto-regressive integrated moving average with X (ARIMAX) use more sophisticated regression models to forecast the future value of a time-series data. It is important to note that using complex mathematical models does not guarantee more accurate forecast. Simple moving average technique may outperform complex ARIMA models in few cases. In fact, in an editorial in the International Journal of Forecasting, Chatfield (1986) claimed that simple forecasting models sometimes outperform complex models.

Usually, many different forecasting techniques such as moving average, exponential smoothing, and ARIMA are used for forecasting before selecting the best model. The model selection may depend on the chosen forecasting accuracy measure. The following four forecasting accuracy measures are frequently used:

1. Mean absolute error
2. Mean absolute percentage error
3. Mean squared error
4. Root mean square error

In the following subsections, we will discuss these measures.

### 13.3.1 | Mean Absolute Error (MAE)

Mean absolute error (MAE) is the average absolute error and should be calculated on the validation data set. Assume that the validation data has  $n$  observations and forecasting is carried out on these  $n$  observations using the model developed. The mean absolute error is given by

$$MAE = \sum_{t=1}^n \frac{|Y_t - F_t|}{n} \quad (13.3)$$

In Eq. (13.3),  $Y_t$  is the actual value of  $Y$  at time  $t$  and  $F_t$  is the corresponding forecasted value.

### 13.3.2 | Mean Absolute Percentage Error (MAPE)

Mean absolute percentage error (MAPE) is the average of absolute percentage error. Assume that the validation data has  $n$  observations and the forecasting is carried out on these  $n$  observations. The mean absolute percentage error is given by

$$MAPE = \frac{1}{n} \sum_{t=1}^n \frac{|Y_t - F_t|}{Y_t} \times 100\% \quad (13.4)$$

MAPE defined in Eq. (13.4) is one of the popular forecasting accuracy measures used by practitioners since it expresses the average error in percentage terms and is easy to interpret. Since MAPE is dimensionless it can be used for comparing different models with varying scales.

### 13.3.3 | Mean Square Error (MSE)

Mean square error is the average of squared error calculated over the validation data set. MSE is given by

$$MSE = \frac{1}{n} \sum_{t=1}^n (Y_t - F_t)^2 \quad (13.5)$$

Lower MSE implies better prediction. However, it depends on the range of the time-series data.

### 13.3.4 | Root Mean Square Error (RMSE)

Root mean square error (RMSE) is the square root of mean square error and is given by

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n (Y_t - F_t)^2} \quad (13.6)$$

RMSE along with MAPE are two most popular accuracy measures of forecasting. RMSE is the standard deviation of errors or residuals. In 2006, Netflix, the movie portal, announced a competition with a prize money worth one million dollars to predict the rating on a 5-point scale likely to be given a customer for a movie<sup>2</sup> (source: Wikipedia). The participants were given a target RMSE of 0.8572 to qualify for the prize.

<sup>2</sup> [https://en.wikipedia.org/wiki/Netflix\\_Prize](https://en.wikipedia.org/wiki/Netflix_Prize)

### 13.4 | MOVING AVERAGE METHOD

Moving average is one of the simplest forecasting techniques which forecasts the future value of a time-series data using average (or weighted average) of the past 'N' observations. Mathematically, a simple moving average is calculated using the formula

$$F_{t+1} = \frac{1}{N} \sum_{k=t+1-N}^t Y_k \quad (13.7)$$

The above formula is called simple moving average (SMA) since 'N' past observations are given equal weights ( $1/N$ ). In a weighted moving average, past observations are given differential weights (usually the weight decrease as the data becomes older). Weighted moving average is given by

$$F_{t+1} = \sum_{k=t+1-N}^t W_k \times Y_k \quad (13.8)$$

where  $W_k$  is the weight given to value of  $Y$  at time  $k$  ( $Y_k$ ) and  $\sum_{k=t+1-N}^t W_k = 1$ .

#### EXAMPLE 13.1

We Sell Beauty (WSB) is a manufacturer and distributor of health and beauty products. WSB is interested in forecasting demand for 'Kesh', their shampoo brand which is sold in 100 ml bottles. WSB believes that the monthly demand for 'Kesh' depends on the promotion expenditure (in thousands of rupees) and whether the competition was on promotion or not during that month. The data for 48 months (starting from January 2012) is shown in Table 13.1. The table has the quantity of 100 ml bottles sold during the month, promotion expenses (in thousands of rupees) incurred by the company, and whether the competition was on promotion (value of 1 implies that the competition was on promotion and 0 otherwise). Use simple moving average with  $N = 12$  and forecast the demand of Kesh for months 37 to 48. Calculate the values of MAPE and RMSE.

**TABLE 13.1** Data on sales of shampoo, promotion expenses (in 1000 of rupees), and dummy variable for promotion by competition

Month	Sale Quantity	Promotion Expenses	Competition Promotion	Month	Sale Quantity	Promotion Expenses	Competition Promotion
1	3002666	105	1	25	4634047	165	0
2	4401553	145	0	26	3772879	129	1
3	3205279	118	1	27	3187110	120	1
4	4245349	130	0	28	3093683	112	1
5	3001940	98	1	29	4557363	162	0

**TABLE 13.1** Data on sales of shampoo, promotion expenses (in 1000 of rupees), and dummy variable for promotion by competition—Continued

Month	Sale Quantity	Promotion Expenses	Competition Promotion	Month	Sale Quantity	Promotion Expenses	Competition Promotion
6	4377766	156	0	30	3816956	140	1
7	2798343	98	1	31	4410887	160	0
8	4303668	144	0	32	3694713	139	0
9	2958185	112	1	33	3822669	141	1
10	3623386	120	0	34	3689286	136	0
11	3279115	125	0	35	3728654	130	1
12	2843766	102	1	36	4732677	168	0
13	4447581	160	0	37	3216483	121	1
14	3675305	130	0	38	3453239	128	0
15	3477156	130	0	39	5431651	170	0
16	3720794	140	0	40	4241851	160	0
17	3834086	167	1	41	3909887	151	1
18	3888913	148	1	42	3216438	120	1
19	3871342	150	1	43	4222005	152	0
20	3679862	129	0	44	3621034	125	0
21	3358242	120	0	45	5162201	170	0
22	3361488	122	0	46	4627177	160	0
23	3670362	135	0	47	4623945	168	0
24	3123966	110	1	48	4599368	166	0

Moving average forecast for the period  $n = 37$  to 48 is given by

$$F_{t+1} = \frac{1}{12} \sum_{k=t+1-12}^t Y_k, \text{ for } t = 36, 37, \dots, 47$$

The forecasted values using 12 period moving average and the corresponding RMSE and MAPE calculations are given in Table 13.2.

**TABLE 13.2** Simple moving average forecast, RMSE, and MAPE calculations

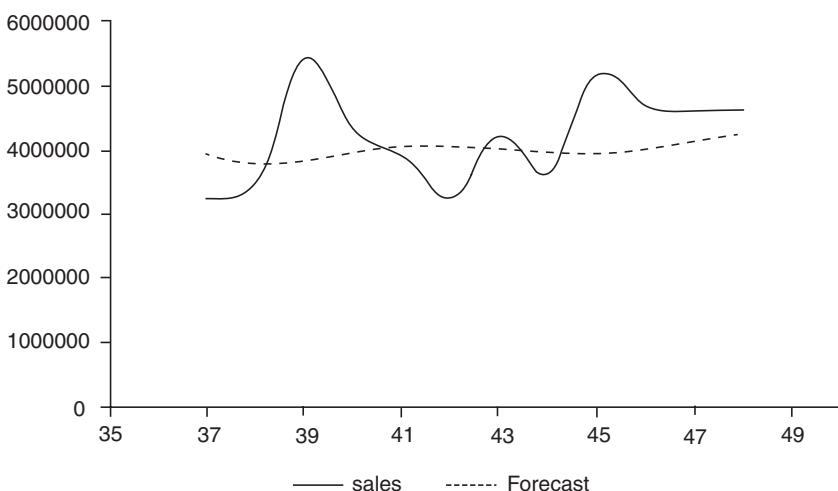
Month	Actual Demand ( $Y_t$ )	Forecasted Demand ( $F_t$ )	$(Y_t - F_t)^2$	$ Y_t - F_t  / Y_t$
37	3216483	3928410	5.07E + 11	0.221337
38	3453239	3810280	1.27E + 11	0.103393
39	5431651	3783643	2.72E + 12	0.303408
40	4241851	3970688	7.35E + 10	0.063926

(Continued)

**TABLE 13.2** Simple moving average forecast, RMSE, and MAPE calculations —Continued

Month	Actual Demand ( $Y_t$ )	Forecasted Demand ( $F_t$ )	$(Y_t - F_t)^2$	$ Y_t - F_t  / Y_t$
41	3909887	4066369	2.45E + 10	0.040022
42	3216438	4012413	6.34E + 11	0.247471
43	4222005	3962370	6.74E + 10	0.061496
44	3621034	3946629	1.06E + 11	0.089918
45	5162201	3940490	1.49E + 12	0.236665
46	4627177	4052117	3.31E + 11	0.124279
47	4623945	4130275	2.44E + 11	0.106764
48	4599368	4204882	1.56E + 11	0.08577

The RMSE using the moving average forecast is given by 734725.8359 and the MAPE value is 0.1403 (or 14.03%). The graph of actual and forecasted demand is shown in Figure 13.2.

**FIGURE 13.2** Plot of actual sales forecasted sales using moving average.

In moving average an important decision that one has to take is the number of periods,  $N$ . The forecast accuracy will depend on the chosen  $N$ . If  $N$  is small, then the average tends to be more sensitive to recent observations or more responsive to recent trend. So, if responsiveness is important, then relatively few data points may be included. This would enable the moving average to make adjustments with the changes in the data quickly, though at times it would also be responding to just the random noise in the data. On the other hand, if  $N$  is large, that is more data points are included, then the forecast will be less sensitive or response to the recent changes in the data. Since the moving average will always be centered around the range of the data points considered, it will lag behind the trend until about  $(N + 1)/2$  time periods.

### 13.5 | SINGLE EXPONENTIAL SMOOTHING (ES)

One of the drawbacks of simple moving average technique is that it gives equal weight to all the previous observations used in forecasting the future value. This can be overcome by assigning differential weights to the past observations [Eq. (13.8)]. One easier way to assign differential weight is achieved by using single exponential smoothing (SES) technique (also known as simple exponential smoothing). Just like the moving average, SES assumes a fairly steady time-series data with no significant trend, seasonal or cyclical component. Here, the weights assigned to past data decline exponentially with the most recent observations assigned higher weights.

In single ES, the forecast at time  $(t + 1)$  is given by (Winters, 1960)

$$F_{t+1} = \alpha Y_t + (1 - \alpha) F_t \quad (13.9)$$

Parameter  $\alpha$  in Eq. (13.9) is called the **smoothing constant** and its value lies between 0 and 1. Since the model uses one smoothing constant, it is called **single exponential smoothing**. Substituting for  $F_t$  recursively in Eq. (13.9), we get

$$F_{t+1} = \alpha Y_t + \alpha(1 - \alpha)Y_{t-1} + \alpha(1 - \alpha)^2 Y_{t-2} + \dots + \alpha(1 - \alpha)^{t-1} Y_1 + (1 - \alpha)^t F_1 \quad (13.10)$$

From Eq. (13.10), it is evident that the weights assigned to older observations decrease exponentially. Figure 13.3 shows the rate at which the weight decreases for older observations when  $\alpha = 0.4$  and  $0.8$ ; the plot resembles the exponential decay curve.

The forecasted values for months 37 to 48 for the data in Table 13.1 using simple exponential smoothing is shown in Table 13.3. Exponential smoothing uses the entire historical data. To begin exponential

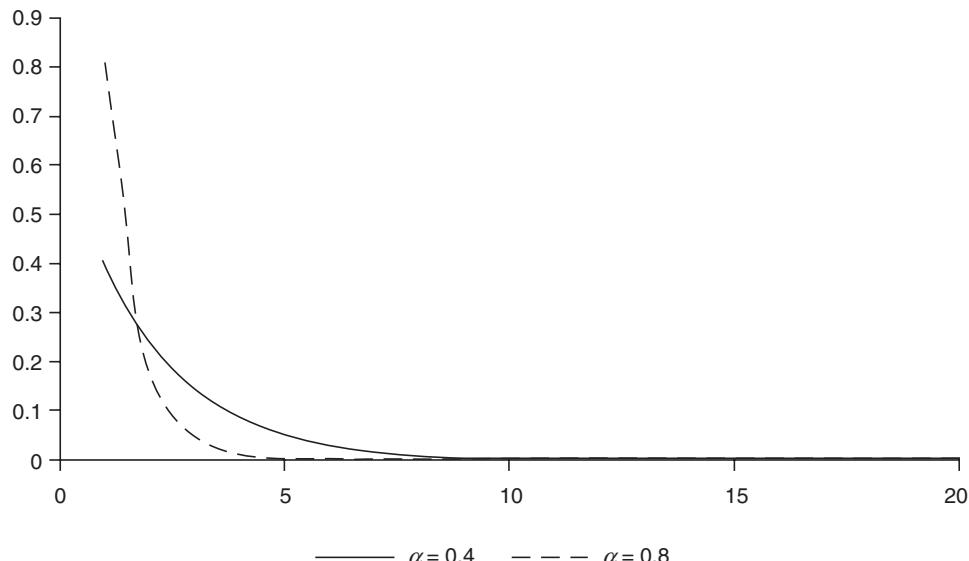


FIGURE 13.3 Exponential decay of weights to older observations.

**TABLE 13.3** Forecast for periods 37 to 48 using single exponential smoothing ( $\alpha = 0.2$ )

Month	Actual Demand	Forecasted Demand $\alpha=0.2$	$(Y_t - F_t)^2$	$\frac{ Y_t - F_t }{Y_t}$
37	3216483	3980905	5.8434E + 11	0.237658
38	3453239	3828020	1.4046E + 11	0.10853
39	5431651	3753064	2.8177E + 12	0.309038
40	4241851	4088781	2.343E + 10	0.036086
41	3909887	4119395	4.3894E + 10	0.053584
42	3216438	4077494	7.4142E + 11	0.267705
43	4222005	3905283	1.0031E + 11	0.075017
44	3621034	3968627	1.2082E + 11	0.095993
45	5162201	3899108	1.5954E + 12	0.244681
46	4627177	4151727	2.2605E + 11	0.102752
47	4623945	4246817	1.4223E + 11	0.08156
48	4599368	4322243	7.6799E + 10	0.060253

smoothing we will need the forecast for the  $F_t$  in Eq. (13.9). We can use  $F_t = Y_t$  or use moving average to forecast the initial forecast  $F_t$ . The forecasted value for period 2 is given by

$$F_2 = \alpha Y_1 + (1 - \alpha) F_1$$

We will assume  $F_1$  same as  $Y_1$ . Thus the value of  $F_2$  will be same as  $Y_1$ , that is 3002666. The forecasted values using single exponential smoothing with  $\alpha = 0.2$  are shown in Table 13.3.

The RMSE using the single exponential smoothing with  $\alpha = 0.2$  is given by 742339.222 and the MAPE value is 0.1394 (or 13.94%).

In summary, single exponential smoothing technique has the following advantages:

1. It uses all the historic data unlike the moving average where only the past few observations are considered to predict the future value.
2. It assigns progressively decreasing weights to older data.

Some disadvantages of smoothing methods are:

1. Increasing  $n$  makes forecast less sensitive to changes in data.
2. It always lags behind trend as it is based on past observations. The longer the time period  $n$ , the greater the lag as it is slow to recognize the shifts in the level of the data points.
3. Forecast bias and systematic errors occur when the observations exhibit strong trend or seasonal patterns.

### 13.5.1 | Optimal Smoothing Constant in a Single Exponential Smoothing (SES)

Choosing optimal smoothing constant  $\alpha$  is important for accurate forecast. Whenever the data is smooth (without much fluctuations), we may choose higher value of  $\alpha$ . However, when the data is highly fluctuating, then it is better to choose lower value of  $\alpha$ . We can find the optimal value of the smoothing constant by solving a non-linear optimization problem. For example, assume that we have to find the optimal  $\alpha$  that will give the minimum RMSE. This can be achieved by solving the following optimization problem:

$$\underset{\alpha}{\text{Min}} \left[ \sqrt{\frac{1}{n} \sum_t (Y_t - F_t)^2} \right] \quad (13.11)$$

subject to the constraint:  $0 < \alpha < 1$ . For the data in Table 13.1, the optimal value of  $\alpha$  that minimizes the RMSE is 0.1574 and the corresponding RMSE is 739399.76. Table 13.4 shows the forecasted value, RMSE, and MAPE calculations for  $\alpha = 0.1574$  (rounded to 4 decimals).

## 13.6 | DOUBLE EXPONENTIAL SMOOTHING – HOLT’S METHOD

One of the drawbacks of single exponential smoothing is that the model does not do well in the presence of trend. This can be improved by introducing an additional equation for capturing the trend in the time-series data. Double exponential smoothing uses two equations to forecast the future values of the time series, one for forecasting the level (short term average value) and another for capturing the trend. The two equations are provided in Eqs. (13.12) and (13.13).

**TABLE 13.4** Forecast for periods 37 to 48 using single exponential smoothing ( $\alpha = 0.1574$ )

Month	Actual Demand	Forecasted Demand $\alpha = 0.1574$	$(Y_t - F_t)^2$	$\frac{ Y_t - F_t }{Y_t}$
37	3216483	3931892	5.1181E + 11	0.22242
38	3453239	3819242	1.3396E + 11	0.105988
39	5431651	3761610	2.789E + 12	0.307465
40	4241851	4024580	4.7207E + 10	0.051221
41	3909887	4058792	2.2173E + 10	0.038084
42	3216438	4035345	6.7061E + 11	0.254601
43	4222005	3906397	9.9608E + 10	0.074753
44	3621034	3956094	1.1227E + 11	0.092532
45	5162201	3903334	1.5847E + 12	0.243862
46	4627177	4101560	2.7627E + 11	0.113594
47	4623945	4184325	1.9327E + 11	0.095075
48	4599368	4253549	1.1959E + 11	0.075188

Level (or Intercept) equation ( $L_t$ ):

$$L_t = \alpha \times Y_t + (1 - \alpha) \times F_{t-1} \quad (13.12)$$

The trend equation is given by ( $T_t$ )

$$T_t = \beta \times (L_t - L_{t-1}) + (1 - \beta) \times T_{t-1} \quad (13.13)$$

$\alpha$  and  $\beta$  are the smoothing constants for level and trend, respectively, and  $0 < \alpha < 1$  and  $0 < \beta < 1$ .

The forecast at time  $t + 1$  is given by

$$F_{t+1} = L_t + T_t \quad (13.14)$$

$$F_{t+n} = L_t + nT_t \quad (13.15)$$

where  $L_t$  is the level which represents the smoothed value up to and including the last data,  $T_t$  is the slope of the line or the rate of increase or decrease at period  $t$ ,  $n$  is the number of time periods into the future.

Initial value of  $L_t$  is usually taken same as  $Y_1$  (that is,  $L_1 = Y_1$ ). The starting value of  $T_t$  can be taken as  $(Y_t - Y_{t-1})$  or the difference between two previous actual values of observations prior to the period for which forecasting is carried out. Another option for  $T_t$  is  $(Y_t - Y_1)/(t - 1)$ .

The value of

$$L_1 = Y_1 = 3002666$$

and

$$T_1 = (Y_{36} - Y_1)/35 = (4732677 - 3002666)/35 = 49428.8857$$

The value of

$$F_2 = L_1 + T_1 = 3002666 + 49428.8857 = 3052095$$

The forecasted values for periods 37 to 48 are shown in Table 13.5 ( $\alpha = 0.0328$  and  $\beta = 0.9486$ ).

The RMSE and MAPE of the forecast using double exponential smoothing is given by 659888.9554 and 0.1135 (11.35%). The values of  $\alpha$  and  $\beta$  used in Table 13.5 are optimized values of  $\alpha$  and  $\beta$  that minimize the root mean square error.

**TABLE 13.5** Forecasted values using double exponential smoothing ( $\alpha = 0.0328$  and  $\beta = 0.9486$ )

Month	Actual Demand	$L_t$	$T_t$	$F_t (= L_{t-1} + T_{t-1})$	$(Y_t - F_t)^2$	$ Y_t - F_t /Y_t$
37	3216483	3678293	66894.6916	3693955	2.27979E+11	0.148445
38	3453239	3735612	57810.9617	3745188	85234318285	0.084544
39	5431651	3847157	108782.913	3793423	2.68379E+12	0.301608
40	4241851	3965318	117678.771	3955940	81745109031	0.067402
41	3909887	4077319	112292.624	4082997	29966946139	0.044275
42	3216438	4157691	82013.2329	4189611	9.47066E+11	0.302562
43	4222005	4239124	81462.532	4239704	313269245.9	0.004192
44	3621034	4297641	59696.6025	4320586	4.89374E+11	0.193191
45	5162201	4383737	84739.1839	4357338	6.47805E+11	0.155915
46	4627177	4473682	89677.0074	4468476	25185883916	0.034298
47	4623945	4565346	91562.092	4563359	3670690475	0.013103
48	4599368	4655021	89771.7849	4656908	3310862728	0.01251

### 13.7 | TRIPLE EXPONENTIAL SMOOTHING (HOLT-WINTER MODEL)

Moving averaging and single and double exponential smoothing techniques discussed so far can handle data as long as the data do not have any seasonal component associated with it. However, when there is seasonality in the time-series data, techniques such as moving average, exponential smoothing, and double exponential smoothing are no longer appropriate. In most cases, the fitted error values (actual demand minus forecast) associated with simple exponential smoothing and Holt's method will indicate systematic error patterns that reflect the existence of seasonality. For example, presence of seasonality may result in all positive errors, except for negative values that occur at fixed intervals. Such pattern in error would imply existence of seasonality. Such time series data require the use of a seasonal method to eliminate the systematic patterns in error.

Triple exponential smoothing is used when the data has trend as well as seasonality. The following three equations which account for level, trend, and seasonality are used for forecasting (for a multiplicative model, Winters 1960):

**Level (or Intercept) equation:**

$$L_t = \alpha \frac{Y_t}{S_{t-c}} + (1-\alpha)[L_{t-1} + T_{t-1}] \quad (13.16)$$

**Trend equation:**

$$T_t = \beta \times (L_t - L_{t-1}) + (1 - \beta) T_{t-1} \quad (13.17)$$

**Seasonal equation:**

$$S_t = \gamma \frac{Y_t}{L_t} + (1 - \gamma) S_{t-c} \quad (13.18)$$

The forecast  $F_{t+1}$  using triple exponential smoothing is given by

$$F_{t+1} = [L_t + T_t] \times S_{t+1-c} \quad (13.19)$$

where  $c$  is the number of seasons (if it is monthly seasonality, then  $c = 12$ ; in case of quarterly seasonality  $c = 4$ ; and in case of daily data  $c = 7$ ). The initial values of  $L_t$  and  $T_t$  are calculated using the following equations:

$$L_t = Y_t \quad (13.20)$$

Alternatively

$$L_t = \frac{1}{c} (Y_1 + Y_2 + \dots + Y_c) \quad (13.21)$$

$$T_t = \frac{1}{c} \left[ \frac{Y_t - Y_{t-c}}{12} + \frac{Y_{t-1} - Y_{t-c-1}}{12} + \frac{Y_{t-2} - Y_{t-c-2}}{12} + \dots + \frac{Y_{t-c+1} - Y_{t-2c+1}}{12} \right] \quad (13.22)$$

Several techniques exist to calculate the initial seasonality index (Winters, 1961; Makridakis *et al.*, 1998; Taylor 2011). The initial seasonality index can be calculated using a technique called method of simple averages which is described in Section 13.7.1. Several variations to the procedure discussed in Section 13.7.1 exist, such as ratio-to-moving average.

### 13.7.1 | Predicting Seasonality Index Using Method of Averages

The following steps are used for predicting the seasonality index using method of averages:

#### STEP 1

Calculate the average of value of  $Y$  for each season (that is, if the data is monthly data, then we need to calculate the average for each month based on the training data). Let these averages be  $\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_c$ .

#### STEP 2

Calculate the average of the seasons' averages calculated in step 1 (say  $\bar{\bar{Y}}$ ).

#### STEP 3

The seasonality index for season  $k$  is given by the ratio  $\bar{Y}_k / \bar{\bar{Y}}$ . Variation to the procedure explained above is first divide the value of  $Y_t$  with its yearly average and calculate the seasonal average. We will use first 3 years of data in Table 13.1 to calculate the seasonality index for various months. The seasonality index based on first 3 years of data using method of averages is shown in Table 13.6.

Seasonality index can be interpreted as percentage change from the trend line. For example, the seasonality index for January is approximately 1.088 or 108.8%. This implies that in January, the demand will be approximately 8.8% more from the trend line. The seasonality index for March is 0.888541 or 88.85%.

**TABLE 13.6** Seasonality index using method of averages

Month	Sale Quantity (2012)	Sale Quantity (2013)	Sale Quantity (2014)	Monthly Average $\bar{Y}_k$	Seasonality Index $\bar{Y}_k / \bar{\bar{Y}}$
January	3002666	4447581	4634047	4028098.00	1.087932
February	4401553	3675305	3772879	3949912.33	1.066815
March	3205279	3477156	3187110	3289848.33	0.888541
April	4245349	3720794	3093683	3686608.67	0.9957
May	3001940	3834086	4557363	3797796.33	1.02573
June	4377766	3888913	3816956	4027878.33	1.087872
July	2798343	3871342	4410887	3693524.00	0.997568
August	4303668	3679862	3694713	3892747.67	1.051375
September	2958185	3358242	3822669	3379698.67	0.912808
October	3623386	3361486	3689286	3558053.33	0.960979
November	3279115	3670362	3728654	3559377.00	0.961337
December	2843766	3123966	4732677	3566803.00	0.963342
Average of monthly averages				3702528.22	

This implies that the demand in March will be 11.15% less from the trend line. Note that, multiplicative model is used in this example.

To start the triple exponential smoothing, we need to set the starting values of level and trend.

$$L_{36} = Y_{36}/S_{36} = 4732677/0.9633 = 4912983.494$$

The initial value of trend ( $T_{36}$ ) can be calculated based on second and third year by using Eq. (13.22):

$$\begin{aligned} T_{36} &= \frac{1}{12} \left[ \frac{Y_{36} - Y_{24}}{12} + \frac{Y_{35} - Y_{23}}{12} + \frac{Y_{34} - Y_{22}}{12} + \dots + \frac{Y_{25} - Y_{13}}{12} \right] \\ T_{36} &= \frac{1}{12} \left[ \frac{4732677 - 3123966}{12} + \frac{3728654 - 3670362}{12} + \dots + \frac{4634047 - 4447581}{12} \right] = 21054.35 \end{aligned}$$

The forecast for period 37 using triple exponential smoothing is given by

$$F_{37} = [L_{36} + T_{36}] \times S_{37-12} = [L_{36} + T_{36}] \times S_{25} \quad (13.23)$$

The seasonal index  $S_{25}$  (seasonality index for January) is 1.088. Substituting the values of  $L_{36}$ ,  $T_{36}$  and  $S_{25}$ , we get

$$F_{37} = [4912983.494 + 21054.35] \times 1.088 = 5368233.2$$

The forecast for the period 37 to 48 for the data in Table 13.1 is given in Table 13.7. Note that the values such as seasonality index are rounded to two decimals, the forecast values will be different if the actual seasonality index values in Table 13.6 are used.

The RMSE and MAPE using triple exponential smoothing are 1228588.29 and 0.2208 (22.08%), respectively. The values of  $\alpha = 0.32$ ,  $\beta = 0.5$ , and  $\gamma = 1$  are used for calculating the level, trend, and seasonal components. It is important to note that the exponential smoothing techniques are very sensitive to initial values of level, trend, and seasonal index.

### 13.8 | CROSTON'S FORECASTING METHOD FOR INTERMITTENT DEMAND

Products such as spare parts may have intermittent demands. Exponential smoothing models discussed so far in the chapter will produce biased estimate when used for intermittent demand. Croston (1972) developed a model that uses two separate exponential smoothing equations for predicting mean time between demands and the magnitude of demand whenever the demand occurs. That is, Croston's method has two components: (a) Predicting time between demand and (b) magnitude of the demand. The primary objective of Croston's method is to forecast mean demand per period. Let

$Y_t$  = Demand at time  $t$  ( $Y_t$  may take value 0)

$F_t$  = Forecasted demand

$TD_t$  = Time between the latest and the previous non-zero demand in period  $t$

$FTD_t$  = Forecasted time between demand at period  $t$

The following steps are used for forecasting demand:

$$\text{If } Y_t = 0 \text{ then } F_{t+1} = F_t \text{ and } FTD_{t+1} = FTD_t \quad (13.24)$$

$$\text{If } Y_t \neq 0 \text{ the } F_{t+1} = \alpha \times Y_t + (1-\alpha)F_t \text{ and } FTD_{t+1} = \beta \times TD_t + (1-\beta) \times FTD_t \quad (13.25)$$

**TABLE 13.7** Forecasting using triple exponential smoothing (values differ for different round off values of parameters)

Month $t$	Actual Demand	$L_{t-1}$	$T_{t-1}$	$S_t$	$F_t$	$(Y_t - F_t)^2$	$ Y_t - F_t /Y_t$
37	3216482	4912983.49	21054.35	1.09	5367895.97	4.62858E+12	0.668872
38	3453239	4301229.28	-295349.93	1.07	4273531.48	6.7288E+11	0.237543
39	5431651	3759825.78	-418376.71	0.89	2969014.38	6.06458E+12	0.453386
40	4241851	4228345.39	25071.45	1.00	4235127.90	45200134.5	0.001585
41	3909887	4255577.53	26151.79	1.03	4391900.21	2.32337E+11	0.123281
42	3216437	4131354.31	-49035.71	1.09	4441041.44	1.49966E+12	0.380733
43	4222004	3722098.55	-229145.74	1.00	3484457.63	5.43975E+11	0.174691
44	3621034	3729543.06	-110850.61	1.05	3804603.81	33697874118	0.050695
45	5162201	3562820.55	-138786.56	0.91	3125486.18	4.14821E+12	0.394544
46	4627176	4138038.05	218215.47	0.96	4186269.09	1.94399E+11	0.095286
47	4623945	4503072.73	291625.07	0.96	4609319.06	213918263.4	0.003163
48	4599368	4799566.34	294059.34	0.96	4906905.01	94579010434	0.066865

$\alpha$  and  $\beta$  are smoothing constants for forecasted demand and forecasted time between demands, respectively. Once the forecasted demand and time between demands are known, then the mean demand per period,  $D_{t+1}$ , is given by

$$D_{t+1} = \frac{F_{t+1}}{FTD_{t+1}} \quad (13.26)$$

**EXAMPLE 13.2**

Quarterly demand for spare parts of avionics system of an aircraft is shown in Table 13.8. Use the demand during the quarters 1 to 4 as training data to forecast demand for periods 5 to 16 using Croston's method.

**TABLE 13.8** Quarterly demand for avionic system spares

Quarter	1	2	3	4	5	6	7	8
Demand	20	12	0	18	16	0	20	22
Quarter	9	10	11	12	13	14	15	16
Demand	0	28	0	0	30	26	0	34

Procedure used for starting values of  $F_t$  and  $FTD_t$  is shown in Table 13.9.

**TABLE 13.9** Calculating initial values in Croston's method

Quarter	Demand	$TD_t$	$FTD_t$	$F_t$
1	20			
2	12	1		
3	0			
4	18	2	1.5	16.67

In Table 13.9,  $TD_4 = 2$  since the elapsed time from the previous demand and current demand period is 2 ( $4 - 2$ ). The forecasted time between demand is the average  $TD_t$  values up to  $t = 4$ . So,  $FTD_4 = (1+2)/2 = 1.5$ . The forecasted demand  $F_4$  for  $t = 4$  is  $(20 + 12 + 18)/3 = 16.67$ . Note that the total value is divided by 3 (not 4) since only 3 quarters had non-zero demand. So, the starting values for Croston's method are

$$TD_4 = 2, FTD_4 = 1.5, \text{ and } F_4 = 16.67$$

Let  $\alpha = \beta = 0.2$ . Then

$$F_5 = 0.2 \times 18 + (1 - 0.2) \times 16.67 = 16.936$$

$$FTD_5 = 0.2 \times 2 + (1 - 0.2) \times 1.5 = 1.6$$

The forecasted values for remaining quarters are shown in Table 13.10.

**TABLE 13.10** Forecasted demand for periods 5 to 16 using Croston's method

Quarter	Demand	$TD_t$	$FTD_t$	$F_t$	$D_t = (F_t/FTD_t)$
1	20				
2	12	1			
3	0				
4	18	2	1.5000	16.67	11.11333
5	16	1	1.6000	16.936	10.585
6	0		1.4800	16.7488	11.31676
7	20	2	1.4800	16.7488	11.31676
8	22	1	1.5840	17.39904	10.98424
9	0		1.4672	18.31923	12.48585
10	28	2	1.4672	18.31923	12.48585
11	0		1.5738	20.25539	12.8707
12	0		1.5738	20.25539	12.8707
13	30	3	1.5738	20.25539	12.8707
14	26	1	1.8590	22.20431	11.94417
15	0		1.6872	22.96345	13.61034
16	34	2	1.6872	22.96345	13.61034

### 13.9 | REGRESSION MODEL FOR FORECASTING

Regression is probably more appropriate method for forecasting when the data has values of predictor (explanatory) variables in addition to the dependent variable  $Y_t$ . In the data provided in Table 13.1, we also have information such as promotion expenses and whether the competition was on promotion or not. Using the values of these predictor variables is likely to give better forecast than the exponential smoothing techniques discussed in the previous sections. Parker and Segura (1971) claimed that regression method can predict more accurately than less scientific methods such as exponential smoothing. The forecasted value at time  $t$ ,  $F_t$ , can be written as a regression equation as follows:

$$F_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \dots + \beta_n X_{nt} + \varepsilon_t \quad (13.27)$$

Here  $F_t$  is the forecasted value of  $Y_t$ , and  $X_{1t}$ ,  $X_{2t}$ , etc. are the predictor variables measured at time  $t$ . The regression equation for Example 13.1 is

$$F_t = \beta_0 + \beta_1 \text{ promotion expenses at time } t + \beta_2 \text{ competition promotion at time } t$$

For the data in Table 13.1, the regression outputs using SPSS are shown in Tables 13.11 and 13.12. The model is developed based on first 36 months data.

**TABLE 13.11** Model summary

Model	R	R-Square	Adjusted R-Square	Std. Error of the Estimate	Durbin–Watson
1	0.928	0.862	0.853	207017.359	1.608

The  $R$ -square for the model is 0.862 (note that we will need high  $R$ -square value for forecasting applications) and the Durbin–Watson statistic value is 1.608. Since this is a time-series data we need to check whether the errors,  $\varepsilon_t$ , are correlated (auto-correlation). For  $n = 36$  (sample size) and number of predictor variables = 2, the Durbin–Watson critical values are  $d_L = 1.153$  and  $d_U = 1.376$ . Since the model Durbin–Watson statistic  $D = 1.608$  ( $4 - D = 2.392$ ) lies within  $d_U$  and  $(4 - d_U)$ , we can conclude that there is no auto-correlation. Whenever regression model is used for forecasting, it should be checked for auto-correlation among the errors. Presence of auto-correlation may lead to inclusion of a non-significant variable in the model since the standard error of the regression coefficient is underestimated when auto-correlation of errors is present.

**TABLE 13.12** Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients		t	Sig.
		B	Std. Error	Beta			
1	(Constant)	808471.843	278944.970			2.898	0.007
	Promotion Expenses	22432.941	1953.674	0.825		11.482	0.000
	Competition Promotion	-212646.036	77012.289	-0.198		-2.761	0.009

The regression model (based on values in Table 13.12) is given by

$$F_t = 808471.843 + 22432.941X_{1t} - 212646.036X_{2t} \quad (13.28)$$

where

$X_{1t}$  = Promotion expenses at time  $t$

$$X_{2t} = \begin{cases} 1 & \text{Competition is on promotion} \\ 0 & \text{Otherwise} \end{cases}$$

As expected, the sales increases as the promotion expenses increase and the sales decreases whenever the competition is on promotion. The forecasted values for period 37 to 48 using the regression model [Eq. (13.28)] is shown in Table 13.13.

TABLE 13.13 Forecasts using regression model

Period	$Y_t$	$X_{1t}$	$X_{2t}$	$F_t$	$(Y_t - F_t)^2$	$\frac{ Y_t - F_t }{Y_t}$
37	3216483	121	1	3310211.67	8785063205	0.02914
38	3453239	128	0	3679888.29	5.137E+10	0.065634
39	5431651	170	0	4622071.81	6.5542E+11	0.149048
40	4241851	160	0	4397742.4	2.4302E+10	0.036751
41	3909887	151	1	3983199.9	5374781013	0.018751
42	3216438	120	1	3287778.73	5089499329	0.02218
43	4222005	152	0	4218278.88	13884007.5	0.000883
44	3621034	125	0	3612589.47	71310120.7	0.002332
45	5162201	170	0	4622071.81	2.9174E+11	0.104632
46	4627177	160	0	4397742.4	5.264E+10	0.049584
47	4623945	168	0	4577205.93	2184540571	0.010108
48	4599368	166	0	4532340.05	4492746215	0.014573

The RMSE and MAPE based on regression model are 302969 and 0.0419 (or 4.19%), respectively. The RMSE and MAPE for regression based forecasting are much smaller than the values that we obtained so far using moving average and exponential smoothing techniques. For Example 13.1, the moving average method resulted in an RMSE of 734725.84 and MAPE is 14.03%. The RMSE and MAPE for single exponential smoothing are 742339.22 and 13.94%, respectively. The plot of actual demand and forecasted demand using regression model is shown in Figure 13.4.

### 13.9.1 | Forecasting Time-Series Data with Seasonal Variation

As mentioned earlier, one can expect seasonal variation in demand for many products and services. The following steps are used to forecast time-series data with seasonal variations:

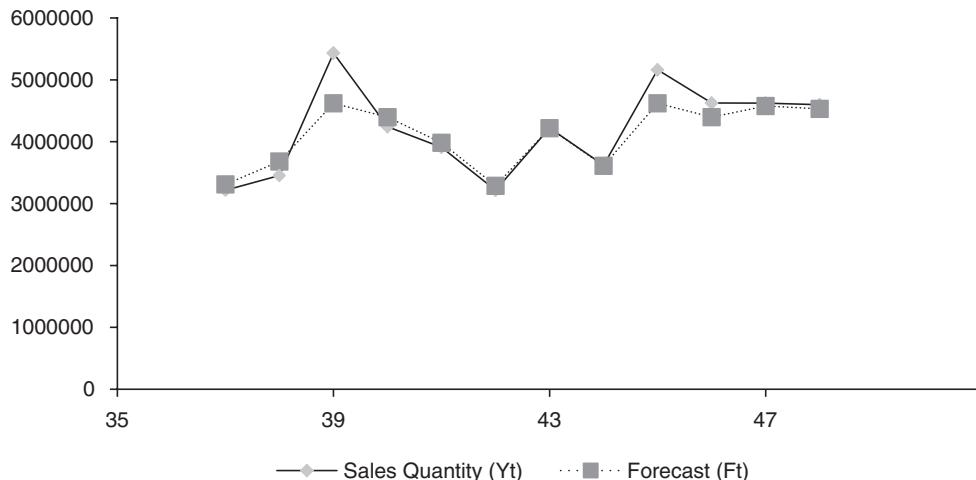


FIGURE 13.4 Actual sales quantity and forecasted sales using regression model.

#### STEP 1

Estimate the seasonality index (using techniques such as method of averages or ratio to moving average).

#### STEP 2

De-seasonalize the data using either additive or multiplicative model. For example, in multiplicative model, the de-seasonalized data  $Y_{d,t} = Y_t / S_t$ , where  $Y_{d,t}$  is the de-seasonalized data and  $S_t$  is the seasonality index for period  $t$ .

#### STEP 3

Develop a forecasting model on the de-seasonalized data ( $F_{d,t}$ ).

#### STEP 4

The forecast for period  $t + 1$  is  $F_{t+1} = F_{d,t+1} \times S_{t+1}$ .

#### EXAMPLE 13.3

Hiccup Viking (HV) is The Vice President of Viking Cookies that specialized into chocolate chip cookies (Choco-Chip). Viking Cookies believes that demand for cookies is seasonal and is driven by several factors such as school holidays, festivals, etc. The shelf life of Choco-Chip cookies is 6 months and excess inventory and

running out of stock can have financial impact. Hiccup would like to develop a forecasting model that they can use for forecasting the demand. The past monthly demand (quantity of 200 gram packets) for four years (January 2013 to December 2016) along with average price per unit during that month is shown in Table 13.14. Develop a forecasting model using regression to predict demand between months 37 and 48, given that the data is seasonal.

**TABLE 13.14** Monthly demand (quantity of 200 gram packets) along with average price per unit

Period	Month	Demand in Units	Average Price	Period	Demand in Units	Average Price
1	January	10500472	37	25	10658309	36
2	February	10123572	34	26	8677622	38
3	March	7372141	36	27	7330354	37
4	April	7764303	38	28	8115471	37
5	May	6904463	40	29	8481936	34
6	June	10068862	34	30	8778999	37
7	July	6436190	40	31	10145039	32
8	August	9898436	34	32	8497839	38
9	September	6803825	39	33	8792138	34
10	October	8333787	36	34	8485358	36
11	November	7541964	39	35	8575904	36
12	December	8540662	37	36	9885156	32
13	January	10229437	37	37	11023467	35
14	February	8453201	38	38	7942451	40
15	March	7997459	35	39	12492798	32
16	April	8557825	35	40	9756258	32
17	May	7818397	36	41	8992741	32
18	June	8944499	37	42	7397807	40
19	July	8904086	36	43	9710611	32
20	August	8463682	39	44	8328379	39
21	September	7723957	37	45	11873063	32
22	October	7731422	39	46	10642507	32
23	November	8441834	35	47	10635075	32
24	December	7485122	40	48	10578547	32

**Solution:**

Since the demand is seasonal, the first step in forecasting is to estimate the seasonality index. We can use first 36 months data to estimate the seasonality index using method of averages explained in Section 13.7.1. Table 13.15 gives the seasonality index for various months. For example, the seasonality index for January is 1.2251. That is, in January the demand will increase by 22.51% from the trend.

**TABLE 13.15** Seasonality index for various months

Month	Demand (2012)	Demand (2013)	Demand (2014)	Average	Seasonality Index
1	10500472	10229437	10658309	10462739	1.2251
2	10123572	8453201	8677622	9084798	1.0637
3	7372141	7997459	7330354	7566651	0.8860
4	7764303	8557825	8115471	8145866	0.9538
5	6904463	7818397	8481936	7734932	0.9057
6	10068862	8944499	8778999	9264120	1.0847
7	6436190	8904086	10145039	8495105	0.9947
8	9898436	8463682	8497839	8953319	1.0483
9	6803825	7723957	8792138	7773307	0.9102
10	8333787	7731422	8485358	8183522	0.9582
11	7541964	8441834	8575904	8186567	0.9585
12	8540662	7485122	9885156	8636980	1.0113
Average of monthly averages				8540659	

De-seasonalized data is calculated by dividing the value of  $Y_t$  with the corresponding seasonality index. The de-seasonalized data for periods 1 to 48 is shown in Table 13.16.

**TABLE 13.16** De-seasonalized demand (seasonality index from Table 13.25 is rounded to 2 decimals)

Month	Demand	Seasonality Index	De-seasonalized Demand	Month	Demand	Seasonality Index	De-seasonalized Demand
1	10500472	1.23	8571459.88	25	10658309	1.23	8700301.09
2	10123572	1.06	9517214.68	26	8677622	1.06	8157870.71
3	7372141	0.89	8321110.54	27	7330354	0.89	8273944.56
4	7764303	0.95	8140603.02	28	8115471	0.95	8508790.52
5	6904463	0.91	7623682.26	29	8481936	0.91	9365476.36
6	10068862	1.08	9282556.42	30	8778999	1.08	8093422.43

**TABLE 13.16** De-seasonalized demand (seasonality index from Table 13.25 is rounded to 2 decimals)—Continued

Month	Demand	Seasonality Index	De-seasonalized Demand	Month	Demand	Seasonality Index	De-seasonalized Demand
7	6436190	0.99	6470703.29	31	10145039	0.99	10199440.54
8	9898436	1.05	9442215.37	32	8497839	1.05	8106172.12
9	6803825	0.91	7475473.63	33	8792138	0.91	9660065.60
10	8333787	0.96	8697481.33	34	8485358	0.96	8855667.03
11	7541964	0.96	7868174.77	35	8575904	0.96	8946835.52
12	8540662	1.01	8445415.13	36	9885156	1.01	9774915.11
13	10229437	1.23	8350215.95	37	11023467	1.23	8998376.94
14	8453201	1.06	7946891.53	38	7942451	1.06	7466733.21
15	7997459	0.89	9026921.81	39	12492798	0.89	14100917.65
16	8557825	0.95	8972583.38	40	9756258	0.95	10229098.91
17	7818397	0.91	8632818.30	41	8992741	0.91	9929490.54
18	8944499	1.08	8245998.07	42	7397807	1.08	6820091.57
19	8904086	0.99	8951833.08	43	9710611	0.99	9762682.98
20	8463682	1.05	8073589.43	44	8328379	1.05	7944522.57
21	7723957	0.91	8486437.69	45	11873063	0.91	13045128.20
22	7731422	0.96	8068828.55	46	10642507	0.96	11106956.05
23	8441834	0.96	8806966.63	47	10635075	0.96	11095071.36
24	7485122	1.01	7401646.68	48	10578547	1.01	10460573.30

Regression output for the de-seasonalized demand and average price using Microsoft Excel are shown in Table 13.17.

**TABLE 13.17** Regression output using SPSS for data in Table 13.16 (based on first 36 cases)

Model	Unstandardized Coefficients		T	Sig.
	B	Std. Error		
1	(Constant)	20812014.673	717702.417	28.998
	Average Price	-335945.859	19616.915	-17.125

Regression model for demand forecasting based on first 36 months of de-seasonalized data is given by

$$F_{d,t} = 20812014.673 - 335945.859 \times \text{Average Price}$$

The forecasted values are given in Table 13.18.

**TABLE 13.18** Forecasted values for the data in Table 13.15

Month	Demand	Seasonality Index ( $S_t$ )	De-seasonalized Demand	$F_{d,t}$	$F_t = F_{d,t} * S_t$	$(Y_t - F_t)^2$	$ Y_t - F_t  / Y_t$
37	11023467	1.2251	8998377	9053910	11091497	4628131462	0.006171
38	7942451	1.0637	7466733	7374180	7844001	9692313467	0.012395
39	12492798	0.8860	14100918	10061747	8914269	$1.2806 \times 10^{13}$	0.286447
40	9756258	0.9538	10229099	10061747	9596642	$2.5477 \times 10^{10}$	0.01636
41	8992741	0.9057	9929491	10061747	9112521	$1.4347 \times 10^{10}$	0.01332
42	7397807	1.0847	6820092	7374180	7998831	$3.6123 \times 10^{11}$	0.081244
43	9710611	0.9947	9762683	10061747	10008080	$8.8488 \times 10^{10}$	0.030633
44	8328379	1.0483	7944523	7710126	8082657	$6.0379 \times 10^{10}$	0.029504
45	11873063	0.9102	13045128	10061747	9157730	$7.373 \times 10^{12}$	0.228697
46	10642507	0.9582	11106956	10061747	9641005	$1.003 \times 10^{12}$	0.094104
47	10635075	0.9585	11095071	10061747	9644592	$9.8106 \times 10^{11}$	0.093134
48	10578547	1.0113	10460573	10061747	10175223	$1.6267 \times 10^{11}$	0.038127

RMSE and MAPE values are 1381119.09 and 0.0775 (7.75%), respectively.

### 13.10 | AUTO-REGRESSIVE (AR), MOVING AVERAGE (MA) AND ARMA MODELS

Auto-regressive (AR) and moving average (MA) models are popular models that are frequently used for forecasting. AR and MA models are combined to create models such as auto-regressive moving average (ARMA) and auto-regressive integrated moving average (ARIMA) models. The initial ARMA and ARIMA models were developed by Box and Jenkins in 1970 (Box and Jenkins, 1970). ARMA models are basically regression models; auto-regression simply means regression of a variable on itself measured at different time periods. One of the fundamental assumptions of AR model is that the time series is assumed to be a stationary process. If a time-series data,  $Y_t$ , is stationary, then it satisfies the following conditions:

1. The mean values of  $Y_t$  at different values of  $t$  are constant.
2. The variances of  $Y_t$  at different time periods are constant (Homoscedasticity).
3. The covariances of  $Y_t$  and  $Y_{t-k}$  for different lags depend only on  $k$  and not on time  $t$ .

When the time series data is not stationary (that is, any one of the above conditions are not satisfied), then we have to convert the non-stationary times-series data to stationary data before applying AR models. Another important concept associated with forecasting based on regression-based models is the white noise of residuals. White noise is a process of residuals  $\varepsilon_t$  that are uncorrelated and follow normal distribution with mean 0 and constant standard deviation. In AR models, one of the important assumptions that we make is that the errors follow a white noise.

### 13.11 | AUTO-REGRESSIVE (AR) MODELS

Auto-regression is regression of a variable on itself measured at different time points. Auto-regressive model with lag 1, AR(1), is given by

$$Y_{t+1} = \beta Y_t + \varepsilon_{t+1} \quad (13.29)$$

which can be re-written as

$$Y_{t+1} - \mu = \beta \times (Y_t - \mu) + \varepsilon_{t+1} \quad (13.30)$$

where  $\varepsilon_{t+1}$  is a sequence of uncorrelated residuals that follow normal distribution with zero mean and constant standard deviation.  $Y_{t+1} - \mu$  can be interpreted as deviation from mean value  $\mu$  and is known as mean centered series. Equation (13.30) can be expanded recursively as shown in Eqs. (13.31) and (13.32):

$$Y_{t+1} - \mu = \beta \times [\beta \times (Y_{t-1} - \mu) + \varepsilon_t] + \varepsilon_{t+1} \quad (13.31)$$

$$Y_{t+1} - \mu = \beta^t (Y_0 - \mu) + \beta^{t-1} \varepsilon_1 + \beta^{t-2} \varepsilon_2 + \dots + \beta \varepsilon_t + \varepsilon_{t+1} \quad (13.32)$$

Equation (13.32) can be written as

$$Y_{t+1} - \mu = \beta^t (Y_0 - \mu) + \sum_{k=1}^{t-1} \beta^{t-k} \times \varepsilon_k + \varepsilon_{t+1} \quad (13.33)$$

In Eq. (13.33), if  $|\beta| > 1$  in the first part on the right-hand side  $[\beta^t (Y_0 - \mu)]$  will result in infinitely large value of  $Y_{t+1}$  as the value of  $t$  increases and is not very useful for practical applications. The value of  $|\beta| = 1$  would imply that the future value of  $Y$  depends on the entire past (and will lead to non-stationarity). For practical applications, the value of  $|\beta|$  should be less than one. The second part of Eq. (13.33),  $\sum_{k=1}^{t-1} \beta^{t-k} \times \varepsilon_k$ , can also become infinitely large if the errors do not follow a white noise. When the errors are white noise then the expected value of  $\sum_{k=1}^{t-1} \beta^{t-k} \times \varepsilon_k$  is zero. The coefficient  $\beta$  in Eq. (13.30) can be estimated using ordinary least squares estimate. The sum of squared errors is given by

$$\sum_{t=2}^n \varepsilon_t^2 = \sum_{t=2}^n [(Y_t - \mu) - \beta \times (Y_{t-1} - \mu)]^2 \quad (13.34)$$

Taking first-derivative of Eq. (13.34) with respect to  $\beta$  and equating that to zero, the estimate of  $\beta$  is given by

$$\hat{\beta} = \frac{\sum_{t=2}^n (Y_t - \mu)(Y_{t-1} - \mu)}{\sum_{t=2}^n (Y_{t-1} - \mu)^2} \quad (13.35)$$

We can generalize auto-regressive model with 1 lag, AR(1), to auto-regressive model with  $p$  lags, AR( $p$ ) process is given by

$$Y_{t+1} = \beta_0 + \beta_1 Y_t + \beta_2 Y_{t-1} + \dots + \beta_p Y_{t-p+1} + \varepsilon_{t+1} \quad (13.36)$$

The forecasted value is given by

$$F_{t+1} = \hat{\beta}_0 + \hat{\beta}_1 Y_t + \hat{\beta}_2 Y_{t-1} + \dots + \hat{\beta}_p Y_{t-p+1} \quad (13.37)$$

where  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots$  are the estimated values  $\beta_0, \beta_1, \beta_2$ , and so on. Note that software such as SPSS use the mean centered series [that is, equation of the form (13.30)] while estimating the parameters and hence forecasting the future values should be carried out using Eq. (13.30).

### 13.11.1 | AR Model Identification: ACF and PACF

One of the important tasks in using auto-regressive model in forecasting is the model identification, which is, identifying the value of  $p$  (the number of lags). One of the standard approaches used for model identification is auto-correlation function (ACF) and partial auto-correlation function (PACF). Auto-correlation is the correlation between  $Y_t$  measured at different time periods (for example,  $Y_t$  and  $Y_{t-1}$  or  $Y_t$  and  $Y_{t-k}$ ). Auto-correlation indicates the memory of a process, that is, how far in time can it remember what has happened before. The auto-correlation of  $k$ -lags (correlation between  $Y_t$  and  $Y_{t-k}$ ) is given by

$$\rho_k = \frac{\sum_{t=k+1}^n (Y_{t-k} - \bar{Y})(Y_t - \bar{Y})}{\sum_{t=1}^n (Y_t - \bar{Y})^2} \quad (13.38)$$

where  $n$  is the number of observations in the sample. A plot of auto-correlation for different values of  $k$  is called **auto-correlation function (ACF)** or **correlogram**.

Partial auto-correlation of lag  $k$  ( $\rho_{pk}$ ) is the correlation between  $Y_t$  and  $Y_{t-k}$  when the influence of all intermediate values ( $Y_{t-1}, Y_{t-2}, \dots, Y_{t-k+1}$ ) is removed (partial out) from both  $Y_t$  and  $Y_{t-k}$ . A plot of partial auto-correlation for different values of  $k$  is called **partial auto-correlation function (PACF)**. Hypothesis tests can be carried out to check whether the auto-correlation and partial auto-correlation values are different from zero. The corresponding null and alternative hypotheses are

$H_0: \rho_k = 0$  and  $H_A: \rho_k \neq 0$ , where  $\rho_k$  is the auto-correlation of order  $k$

$H_0: \rho_{pk} = 0$  and  $H_A: \rho_{pk} \neq 0$ , where  $\rho_{pk}$  is the partial auto-correlation of order  $k$

The null hypothesis is rejected when  $|\rho_k| > 1.96/\sqrt{n}$  and  $|\rho_{pk}| > 1.96/\sqrt{n}$ . To select the appropriate  $p$  in the auto-regressive model, the following thumb rule may be used. The number of lags is  $p$  when (Yaffee and McGee, 2000)

1. The partial auto-correlation,  $|\rho_{pk}| > 1.96 / \sqrt{n}$  for first  $p$  values (first  $p$  lags) and cuts off to zero.
2. The auto-correlation function (ACF),  $\rho_k$ , decreases exponentially.

Note that the model identification is an iterative process and may require additional inputs. The model identification using ACF and PACF cannot be taken as conclusive evidence for the number of lags in AR process.

**EXAMPLE 13.4**

Dr Dawai Sundari (DS) is concerned about the amount of food wasted at her Die Another Day (DAD) hospital. DAD hospital prepares food for all their patients which accounted for 4% of the operating cost. Dr DS found that as high as 50% of the food prepared was wasted on few days. Dr DS believed that an accurate forecasting model will help her reduce the food wastage. The demand for continental breakfast at DAD hospital for past 37 days is shown in Table 13.19. Build an auto-regressive model based on the first 30 days of data and forecast the demand for continental breakfast on days 31 to 37. Comment on the accuracy of the forecast.

**TABLE 13.19** Demand for continental breakfast at DAD

Day	Demand CB	Day	Demand CB
1	25	20	43
2	25	21	41
3	25	22	46
4	35	23	41
5	41	24	40
6	30	25	32
7	40	26	41
8	40	27	41
9	40	28	40
10	40	29	43
11	40	30	46
12	40	31	45
13	44	32	45
14	49	33	46
15	50	34	43
16	45	35	40
17	40	36	41
18	42	37	41
19	40		

**Solution:**

The first step in AR model building is the identification of the right value of  $p$  using ACF and PACF plots. ACF and PACF based on the first 30 observations are given in Figures 13.5 and 13.6, respectively. The horizontal lines in the plot represent the upper and lower critical values for  $\rho_k$  and  $\rho_{pk}$ . The correlation values (vertical bars) beyond the critical values will result in rejection of the null hypothesis.

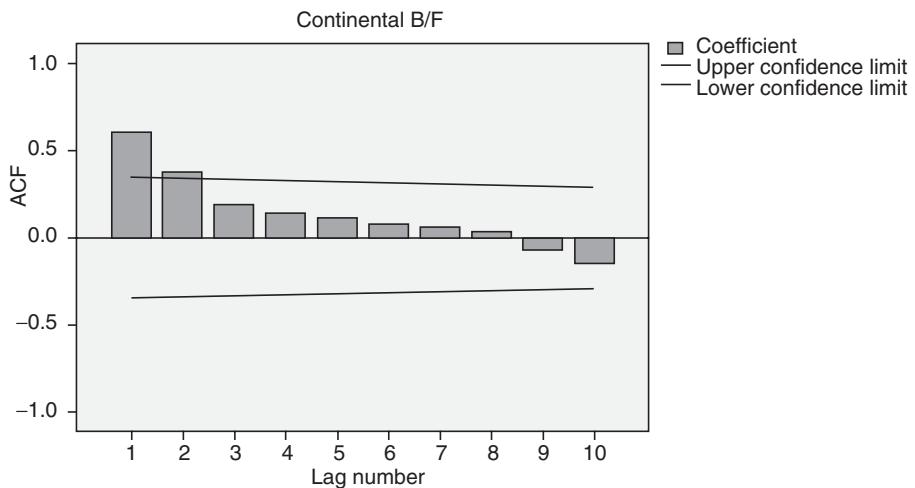


FIGURE 13.5 ACF plot for demand for continental breakfast.

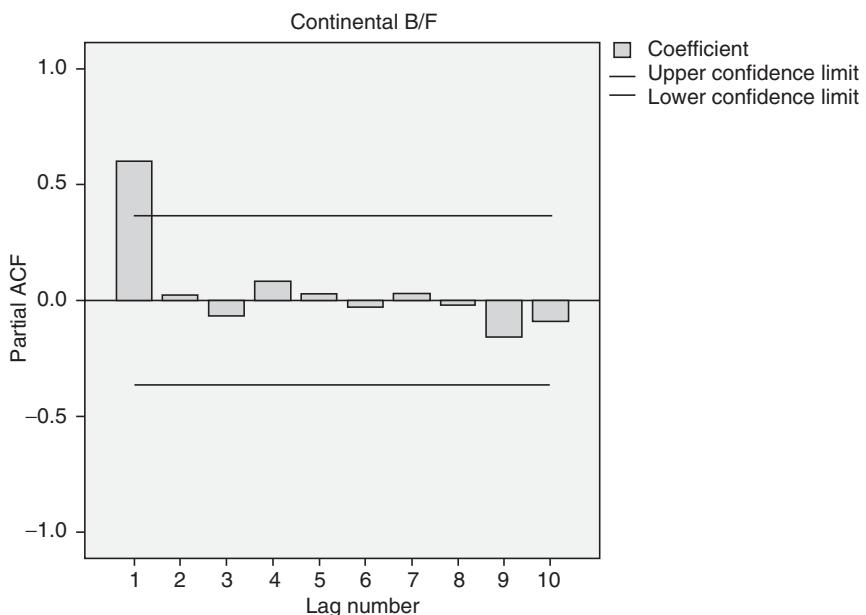


FIGURE 13.6 PACF plot for demand for continental breakfast.

In Figure 13.6, the PACF values cut-off to zero after lag 1 and in Figure 13.5 ACF, the values of auto-correlations are decreasing exponentially. Thus, we can conclude that the value of  $p$  in this case is 1 (note that this is a thumb rule, the correct model may be different from what we identify using ACF and PACF plots). Presence of outliers may have series impact on ACF and PACF and thus the identification of lag value. The values of  $R^2$ , RMSE, MAPE, and regression parameter estimates of AR(1), using SPSS are shown in Tables 13.20 and 13.21.

**TABLE 13.20** AR(1) model statistics

Model	Model Fit Statistics			
	R-Square	RMSE	MAPE	Normalized BIC
Continental B/F-Model_1	0.373	5.133	10.518	3.498

**TABLE 13.21** ARIMA model parameters

			Estimate	SE	T	Sig.
Continental B/F-Model_1	Continental B/F	Constant	38.890	2.995	12.985	0.000
		AR Lag 1	0.731	0.130	5.616	0.000

The residual ACF and PACF plots are shown in Figure 13.7. It is important that the residual ACF and PACF follow a white noise. Since all the auto- (and partial) correlation values are within the critical values, we can conclude that residuals are uncorrelated. The normal P-P plot is shown in Figure 13.8 showing approximate normal distribution for residuals. Since the residual follows white noise as evident from Figures 13.7 and 13.8, we can accept the AR(1) model. Also, the  $p$ -value for lag 1 term in Table 13.21 is less than 0.05. The AR(1) model based on Table 13.21 is given by

$$(F_{t+1} - 38.890) = 0.731(Y_t - 38.890) \quad (13.39)$$

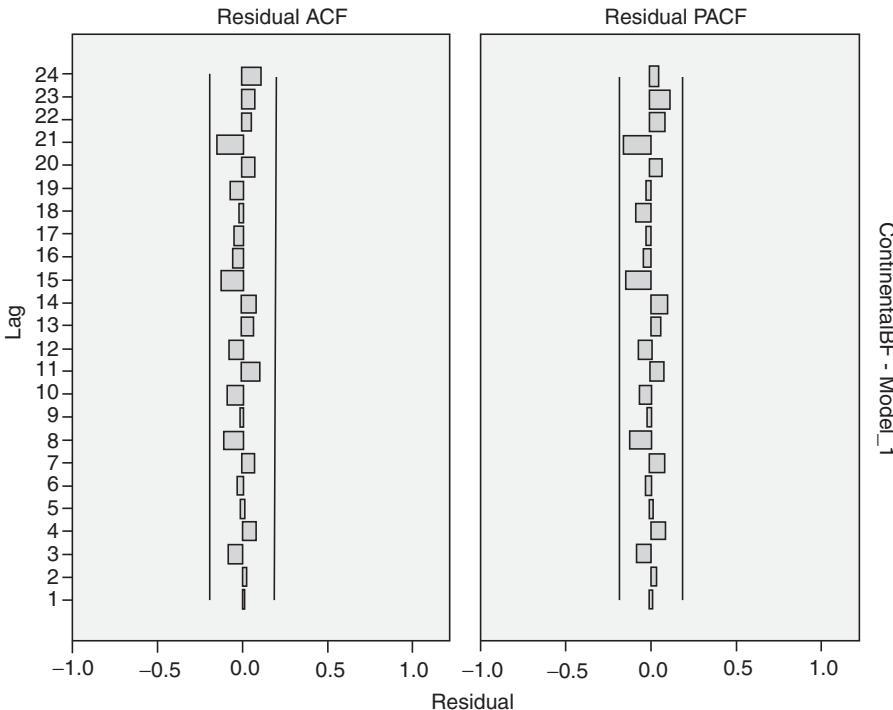
The following equation can be used when we need to forecast several periods in future using forecasted values repetitively.

$$(F_{t+k} - 38.890) = 0.731(F_{t+k-1} - 38.890) \quad (13.40)$$

Using Eq. (13.39), the forecasted value for  $F_{31}$  is given by

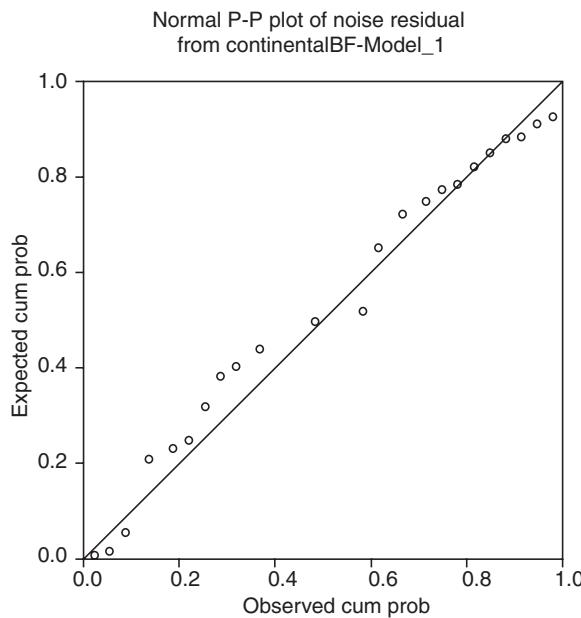
$$(F_{31} - 38.890) = 0.731(Y_{30} - 38.890)$$

$$\Rightarrow F_{31} = 38.890 + 0.731(46 - 38.890) = 44.08$$



**FIGURE 13.7** ACF and PACF plot of residuals.

The forecasted values for periods 31–37 and RMSE and MAPE calculations are shown in Table 13.22.



**FIGURE 13.8** Normal P-P plot of residuals.

**TABLE 13.22** AR(1) model forecast

Day	$Y_t$	$F_t$	$(Y_t - F_t)^2$	$ Y_t - F_t /Y_t$
31	45	44.08741	0.832821	0.02028
32	45	43.35641	2.701388	0.036524
33	46	43.35641	6.988568	0.057469
34	43	44.08741	1.182461	0.025289
35	40	41.89441	3.588789	0.04736
36	41	39.70141	1.686336	0.031673
37	41	40.43241	0.322158	0.013844

The RMSE and MAPE for the validation data (days 31 and 37) are 1.5721 and 0.0332 (3.32%), respectively. Applying Eq. (13.40) that is, using the  $F_t$  values to forecast  $F_{t+1}$  instead of  $Y_t$ , we get the values in Table 13.23.

**TABLE 13.23** AR(1) model forecast using equation 13.40

Day	$Y_t$	$F_t$	$(Y_t - F_t)^2$	$ Y_t - F_t /Y_t$
31	45	44.0874	0.8328	0.0203
32	45	42.6893	5.3393	0.0513
33	46	41.6673	18.7723	0.0942
34	43	40.9202	4.3256	0.0484
35	40	40.3741	0.1399	0.0094
36	41	39.9749	1.0509	0.0250
37	41	39.6830	1.7344	0.0321

The RMSE and MAPE for the validation data (days 31 to 37) using Eq. (13.40) are 2.1446 and 0.04009 (4.009%), respectively.

### 13.12 | MOVING AVERAGE PROCESS MA( $q$ )

Moving average (MA) processes are regression models in which the past residuals are used for forecasting future values of the time-series data. Moving average process is different from moving average technique discussed in Section 13.4, except that the regression model of MA process can be considered as weighted moving average of past residuals. Moving average process of lag 1, MA(1), is given by

$$Y_{t+1} = \mu + \alpha_1 \varepsilon_t + \varepsilon_{t+1} \quad (13.41)$$

Alternatively, a moving average process of lag 1 can be written as

$$Y_{t+1} = \alpha_1 \varepsilon_t + \varepsilon_{t+1} \quad (13.42)$$

MA(1) process uses the previous residual,  $\varepsilon_t$ , to forecast the next value of the time series. The reasoning behind MA process is that the error (also called shock or innovation) at the current period,  $\varepsilon_t$ , and the error at the next period,  $\varepsilon_{t+1}$ , drive the next value of the time series  $Y_{t+1}$ . A moving average process with  $q$  lags, MA( $q$ ) process, is given by

$$Y_{t+1} = \mu + \alpha_1 \varepsilon_t + \alpha_2 \varepsilon_{t-1} + \dots + \alpha_q \varepsilon_{t-q+1} + \varepsilon_{t+1} \quad (13.43)$$

The value of  $q$  (number of lags) in a moving average process can be identified using the following rule (Yaffee and McGee, 2000):

1. Auto-correlation value,  $|\rho_p| > 1.96 / \sqrt{n}$  for first  $q$  values (first  $q$  lags) and cuts off to zero.
2. The partial auto-correlation function,  $\rho_{pk}$ , decreases exponentially.

### 13.13 | AUTO-REGRESSIVE MOVING AVERAGE (ARMA) PROCESS

Auto-regressive moving average (ARMA) is a combination auto-regressive and moving average process. ARMA( $p, q$ ) process combines AR( $p$ ) and MA( $q$ ) processes. ARMA( $p, q$ ) model is given by

$$Y_{t+1} = \underbrace{\beta_1 Y_t + \beta_2 Y_{t-1} + \dots + \beta_p Y_{t-p+1}}_{\text{Auto Regressive Part}} + \underbrace{\alpha_1 \varepsilon_t + \alpha_2 \varepsilon_{t-1} + \dots + \alpha_q \varepsilon_{t-q+1}}_{\text{Moving Average Part}} + \varepsilon_{t+1} \quad (13.44)$$

The parameters are estimated using Box–Jenkins methodology. The values of  $p$  and  $q$  in a ARMA process can be identified using the following thumb rule:

1. Auto-correlation value,  $|\rho_p| > 1.96 / \sqrt{n}$  for first  $q$  values (first  $q$  lags) and cuts off to zero.
2. Partial auto-correlation function,  $|\rho_{pk}| > 1.96 / \sqrt{n}$  for first  $p$  values and cuts off to zero.

#### EXAMPLE 13.5

Monthly demand for avionic system spares used in Vimana 007 aircraft is provided in Table 13.24. Build an ARMA model based on the first 30 months of data and forecast the demand for spares for months 31 to 37. Comment on the accuracy of the forecast.

**TABLE 13.24** Demand for avionic spare parts

Month	Demand for Spares	Month	Demand for Spares
1	457	20	516
2	439	21	656
3	404	22	558
4	392	23	647
5	403	24	864
6	371	25	610
7	382	26	677
8	358	27	609
9	594	28	673
10	482	29	400
11	574	30	443
12	704	31	503
13	486	32	688
14	509	33	602
15	537	34	629
16	407	35	823
17	523	36	671
18	363	37	487
19	479		

**Solution:**

The first step in building ARMA model is the model identification using ACF and PACF plots. The ACF and PACF plots based on the first 30 months demand data are given in Figures 13.9 and 13.10.

In Figure 13.9, the auto-correlations cuts off to zero after 2 lags (note that auto-correlation of lag 3 is just below the critical value). The PACF value cuts off to zero after the first lag. So, the appropriate model could be ARMA(1, 2) process. The SPSS output for ARMA(1, 2) process is given in Tables 13.25 and 13.26.

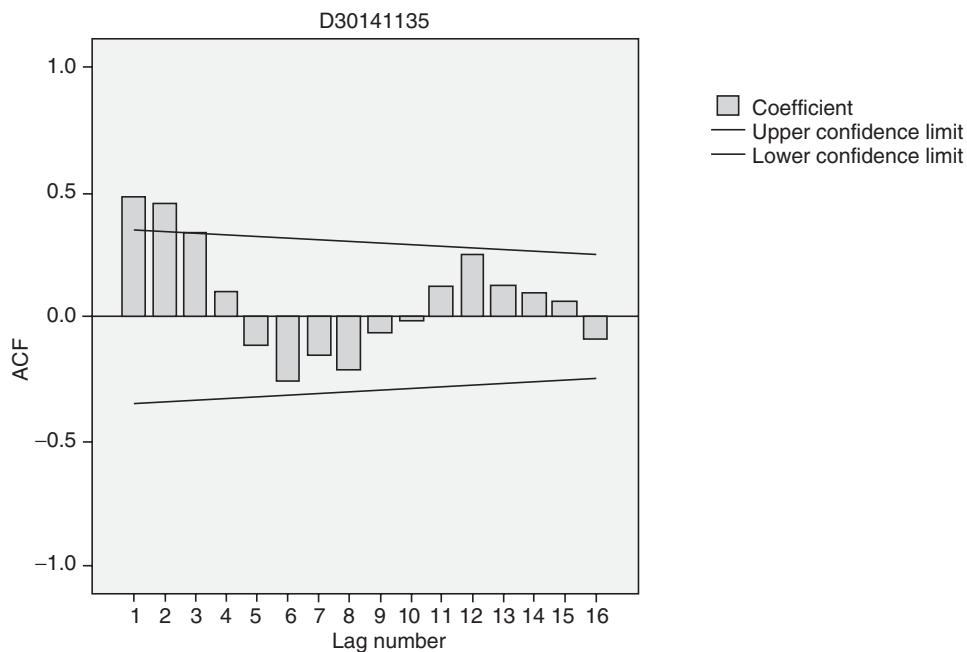


FIGURE 13.9 ACF plot for avionic system spares demand.

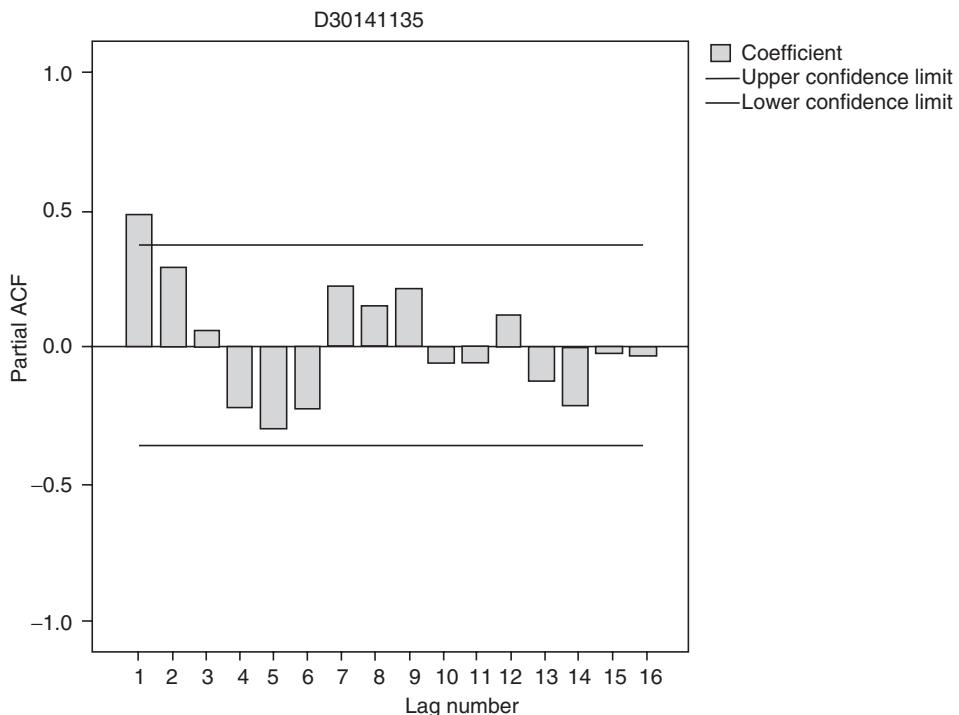


FIGURE 13.10 PACF plot for avionic system spares demand.

**TABLE 13.25** Summary statistics

Model	Model Fit Statistics		
	Stationary R-Squared	RMSE	MAPE
Avionic Spares	0.429	98.824	14.231

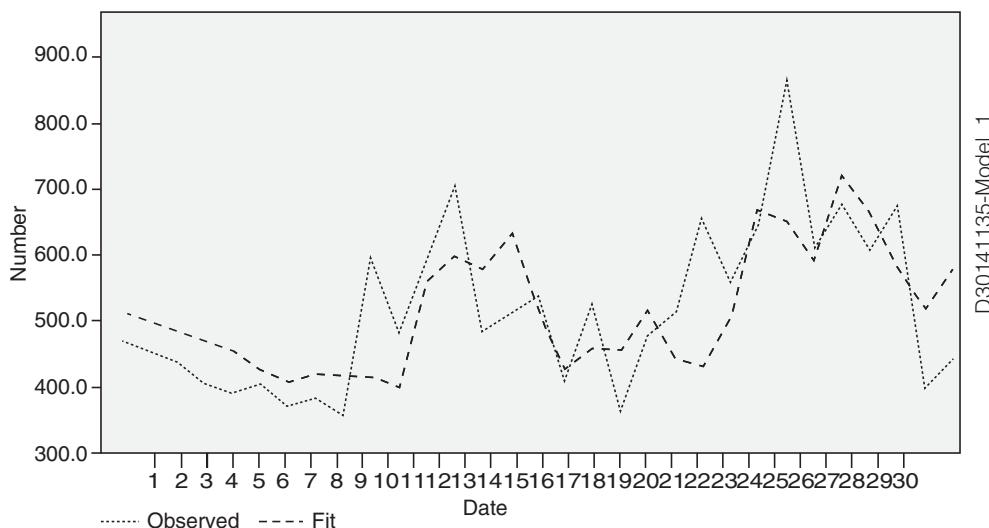
**TABLE 13.26** ARIMA model parameters

			Estimate	SE	T	Sig.
Avionic Spares	Constant		496.699	57.735	8.603	0.000
	AR	Lag 1	0.706	0.170	4.153	0.000
	MA		0.694	0.173	4.006	0.000
	Lag 2		-0.727	0.170	-4.281	0.000

All the three components in the ARMA model (AR lag 1 and MA lags 1 and 2) are statistically significant (Table 13.26). The model equation using SPSS is given by

$$Y_{t+1} - 496.669 = 0.706 \times (Y_t - 496.699) - 0.694 \times \epsilon_t + 0.727 \times \epsilon_{t-1} \quad (13.45)$$

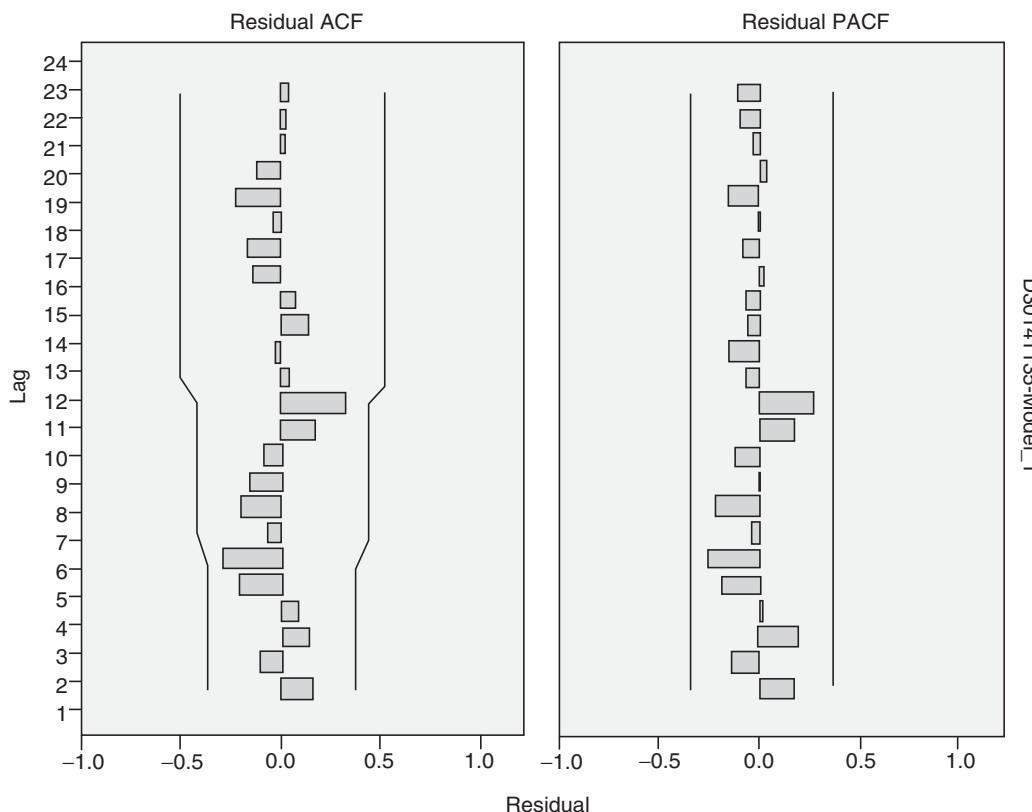
The plot of forecasted and actual values for demand for avionic spare parts based on the model in Eq. (13.45) is given in Figure 13.11. The plot of residual ACF and PACF is shown in Figure 13.12 (it is evident that the residuals are white noise).

**FIGURE 13.11** Observed versus forecasted demand.

**TABLE 13.27** ARMA(1, 2) model forecast

Month	$Y_t$	$F_t$	$(Y_t - F_t)^2$	$ Y_t - F_t /Y_t$
31	503	464.8107	1458.423	0.075923
32	688	378.5341	95769.15	0.449805
33	602	444.6372	24763.04	0.2614
34	629	685.8851	3235.909	0.090437
35	823	743.5124	6318.281	0.096583
36	671	630.7183	1622.614	0.060032
37	487	649.3491	26357.22	0.333366

The RMSE and MAPE for the validation data (months 31 and 37) are 150.961 and 0.1953 (19.53%), respectively (Table 13.27).

**FIGURE 13.12** ACF and PACF of residuals.

The forecasted values using  $F_t$  instead of  $Y_t$  when forecasting for more than one period ahead in time are shown in Table 13.28.

**TABLE 13.28** ARMA (1, 2) forecast

Month	$Y_t$	$F_t$	$(Y_t - F_t)^2$	$ Y_t - F_t /Y_t$
31	503	464.4239	1488.1147	0.0767
32	688	377.8374	96200.8258	0.4508
33	602	444.5195	24800.1101	0.2616
34	629	687.2082	3388.1980	0.0925
35	823	744.9583	6090.4998	0.0948
36	671	630.5592	1635.4571	0.0603
37	487	648.3959	26048.6313	0.3314

The RMSE and MAPE for the validation data (months 31 and 37) are 151.02 and 0.1954 (19.54%), respectively.

### 13.14 | AUTO-REGRESSIVE INTEGRATED MOVING AVERAGE (ARIMA) PROCESS

ARMA models can be used only when the time-series data is stationary. ARIMA models are used when the time-series data is non-stationary. ARIMA model was proposed by Box and Jenkins (1970) and thus is also known as Box–Jenkins methodology. ARIMA has the following three components and is represented as ARIMA( $p, d, q$ ):

1. Auto-regressive component with  $p$  lags AR( $p$ ).
2. Integration component ( $d$ ).
3. Moving average with  $q$  lags, MA( $q$ ).

The main objective of integration component is to convert a non-stationary time-series process to stationary process so that AR and MA processes can be used for forecasting. When the data is non-stationary, the auto-correlation function will not be cut-off to zero quickly; rather ACF may show a very slow decline. When the time series is non-stationary, the model parameter values will be greater than or equal to one resulting in non-convergence of the series and this causes a slow decrease in the values of ACF and PACF. Thus, the presence of non-stationarity can be identified by plotting ACF. A slow decrease in ACF can be diagnosed as non-stationary process. Sample ACF plot for non-stationary data is shown in Figure 13.13.

The non-stationarity could arise from deterministic or stochastic trend; identification of the source of non-stationarity will be useful for using appropriate transformation to make the process stationary. In addition to the visual evidence from ACF plot, Dickey–Fuller or augmented Dickey–Fuller tests are used to check the presence of stationarity.

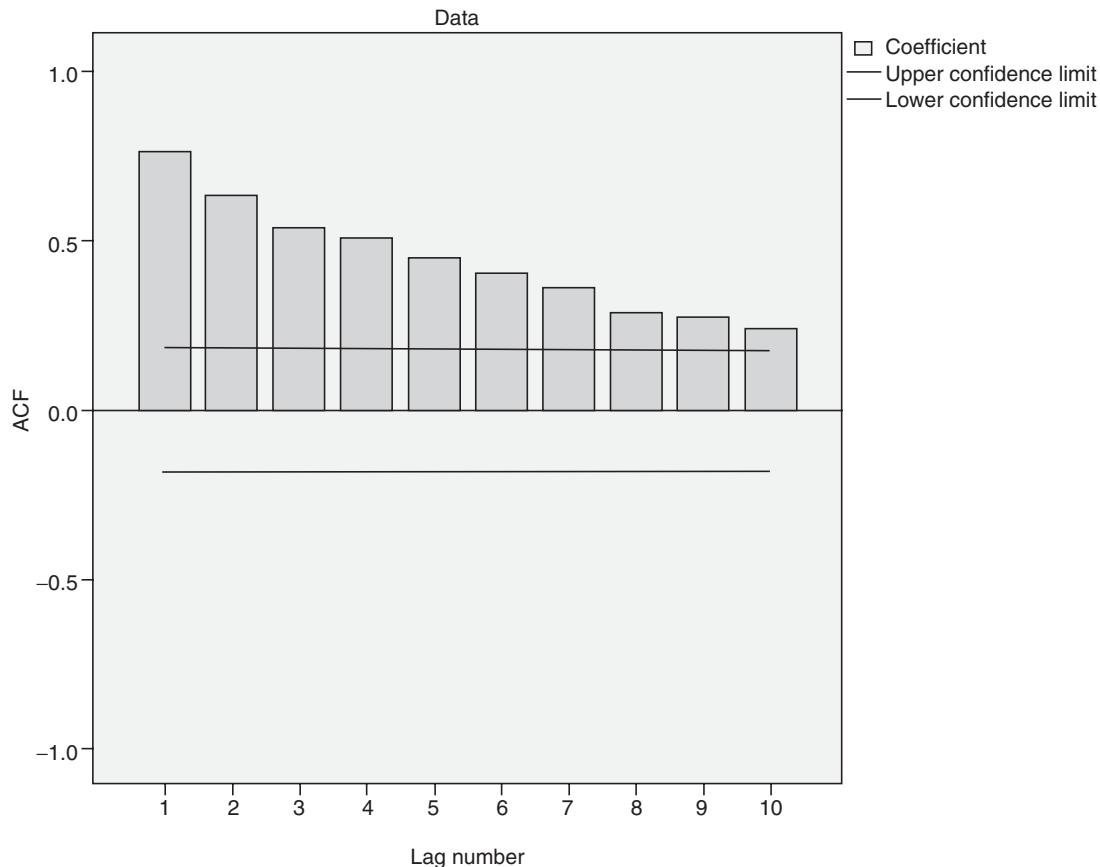


FIGURE 13.13 ACF of a non-stationary process (slowly decreasing auto-correlation values).

### 13.14.1 | Dickey Fuller Test

Consider AR(1) process defined below:

$$Y_{t+1} = \beta Y_t + \varepsilon_{t+1}$$

In Section 13.11, we proved that the AR(1) process can become very large when  $\beta > 1$  and is non-stationary when  $|\beta| = 1$ . Dickey–Fuller test (Dickey and Fuller, 1979) is a hypothesis test in which the null hypothesis and alternative hypothesis are given by

$$H_0: \beta = 1 \text{ (the time series is non-stationary)}$$

$$H_1: \beta < 1 \text{ (the time series is stationary)}$$

The AR(1) can be written as

$$Y_{t+1} - Y_t = \Delta Y_t = (\beta - 1)Y_t + \varepsilon_{t+1} = \psi Y_t + \varepsilon_{t+1} \quad (13.46)$$

In Eq. (13.46),  $\psi = 0$  is same as  $\beta = 1$ . So, the Dickey–Fuller test can be written in terms of  $\psi$  as

$$H_0: \psi = 0 \text{ (the time series is non-stationary)}$$

$$H_A: \psi < 0 \text{ (the time series is stationary)}$$

The test statistic is given by

$$\text{DF Test Statistic} = \frac{\psi}{S_e(\psi)} \quad (13.47)$$

where  $S_e$  is the standard error. Note that DF test statistic is not  $t$ -statistic since the null hypothesis is on non-stationary process. Critical values are derived based on simulation (Fuller, 1976).

### 13.14.2 | Augmented Dickey–Fuller Test

Dickey–Fuller test is valid only when the residual  $\varepsilon_{t+1}$  follows a white noise. When  $\varepsilon_{t+1}$  is not white noise, the actual series may not be AR(1); it may have more significant lags. To address this issue, we augment  $p$ -lags of the dependent variable  $Y$ . The model in Eq. (13.46) can be written as

$$\Delta Y_t = \psi Y_t + \sum_{i=0}^p \alpha_i \Delta Y_{t-i} + \varepsilon_{t+1} \quad (13.48)$$

The above equation can be now tested for non-stationarity. Again the null and alternative hypotheses are

$$H_0: \psi = 0 \text{ (the time series is non-stationary)}$$

$$H_A: \psi < 0 \text{ (the time series is stationary)}$$

### 13.14.3 | Transforming Non-Stationary Process to Stationary Process Using Differencing

The first step in ARIMA is to identify the order of differencing ( $d$ ) required to convert a non-stationary process into a stationary process. Many time-series data will be non-stationary due to factors such as trend and seasonality. If the non-stationary behaviour is due to trend, then it can be converted into a stationary process by de-trending the data. De-trending is usually achieved by fitting a trend line and subtracting it from the time series; this is known as **trend stationarity**. When the reason is not due to trend stationarity, then differencing the original time series may be useful for converting the non-stationary process into a stationary process (called **difference stationarity**).

The first difference ( $d = 1$ ) is the difference between consecutive values of the time series ( $Y_t$  and  $Y_{t-1}$ ). That is, the first difference  $\Delta Y_t$  is given by

$$\nabla Y_t = Y_t - Y_{t-1} \quad (13.49)$$

The second difference ( $d = 2$ ) is the difference of the first differences and is given by

$$\nabla^2 Y_t = \nabla(\nabla Y_t) = Y_t - 2Y_{t-1} + Y_{t-2} \quad (13.50)$$

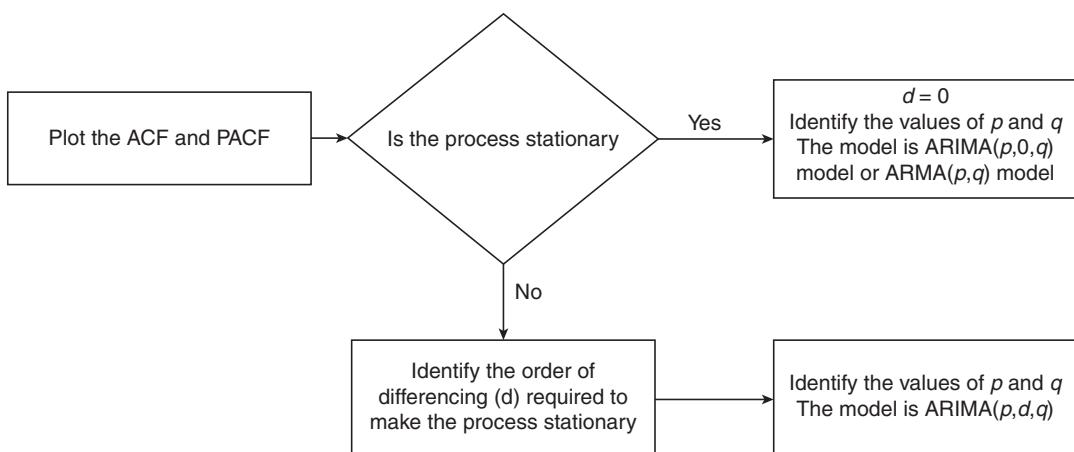
In most cases,  $d \leq 2$  will be sufficient to convert a non-stationary process to a stationary process.

### 13.14.4 | ARIMA( $p, d, q$ ) Model Building

The first step in ARIMA( $p, d, q$ ) is the model identification, that is, identifying the values of  $p$ ,  $d$ , and  $q$ . Box and Jenkins (1970) proposed the following procedure to build the ARIMA( $p, d, q$ ) model.

#### STAGE 1 Model Identification

The main objective of model identification stage is to identify the right values of  $p$  (auto-regressive lags),  $d$  (order of differencing), and  $q$  (moving average lags). The following flow chart can be used during the model identification stage (Figure 13.14). The first step is to plot the ACF and PACF to identify whether the time series is stationary or not. If the time series is stationary then  $d = 0$  and the model is ARIMA( $p, 0, q$ ) or ARMA( $p, q$ ) model. If the time series is non-stationary then it has to be converted into a stationary process by identifying the order of differencing. Once the value of  $d$  is known that will make the process stationary, then  $p$  and  $q$  are identified for the stationary process.



**FIGURE 13.14** Model identification in ARIMA model.

#### STAGE 2 Parameter Estimation and Model Selection

Once the model is identified (values of  $p$ ,  $d$ , and  $q$ ), the next step in ARIMA model building is the parameter estimation. That is, the estimation of coefficients in AR and MA components which are achieved using ordinary least squares. The model selection may be carried using several criteria such as RMSE, MAPE, Akaike Information Criteria (AIC), or Bayesian Information Criteria (BIC).

AIC and BIC are measures of distance from the actual values to the forecasted values. AIC is given by

$$\text{AIC} = -2LL + 2K \quad (13.51)$$

where  $LL$  is the log likelihood function and  $K$  is the number of parameters estimated (in this case  $p + q$ ). BIC is given by

$$\text{BIC} = -2LL + K \ln(n) \quad (13.52)$$

In BIC equation,  $n$  is the number of observations in the sample. BIC assigns higher penalty compared to AIC for every additional variable added to the model. Lower values of AIC and BIC are preferred.

### STAGE 3 Model Validation

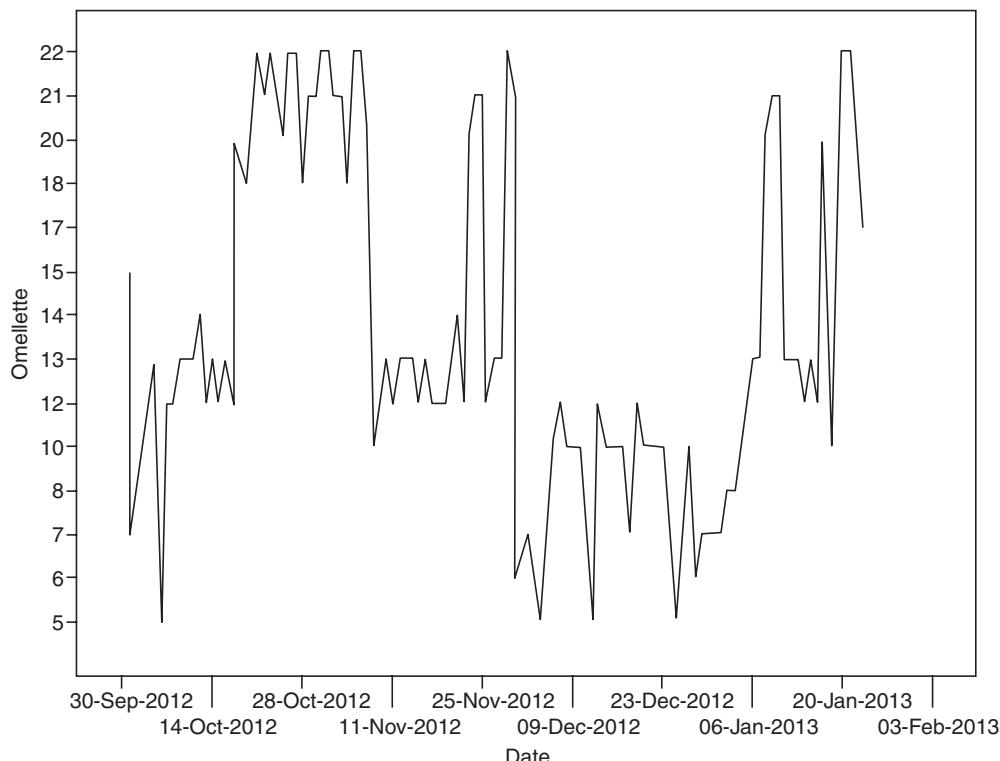
ARIMA model is a regression model and thus has to satisfy all the assumptions of regression. The residual should be white noise. We can also perform a goodness of fit test using Ljung–Box test before accepting the model.

#### EXAMPLE 13.6

Daily demand for Omelette at Die Another Day (DAD) hospital for the past 115 days is given in the excel sheet Example 13.6.xlsx. Develop an appropriate ARIMA model that DAD hospital can use for forecasting demand for Omelette.

#### Solution:

The time-series plot of the daily demand for Omelette is shown in Figure 13.15. The corresponding ACF plot is shown in Figure 13.16. From Figure 13.15, it is evident that the mean is not constant for different values of  $t$ .



**FIGURE 13.15** Time-series plot of demand for Omelette at DAD hospital.

Since the ACF plot shows a very slowly decreasing pattern, we may conclude that the time series is not stationary. We have to convert the process to a stationary process before we can develop a forecasting model. The ACF and PACF plots after differencing ( $d = 1$ ) are shown in Figures 13.17 and 13.18, respectively.

Since both ACF and PACF values are cutting off to zero after the first difference, we may conclude that the appropriate model is ARIMA(1,1,1). Note that subsequent

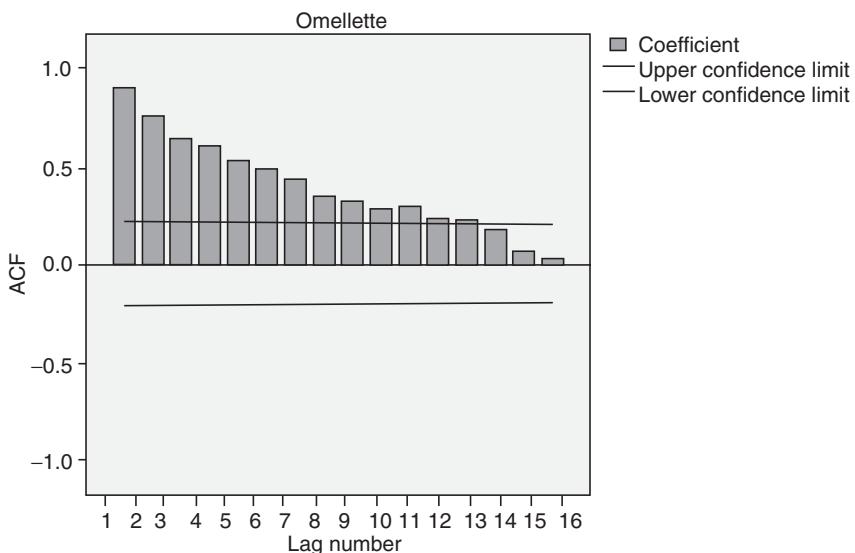


FIGURE 13.16 ACF plot of demand for Omelette at DAD hospital.

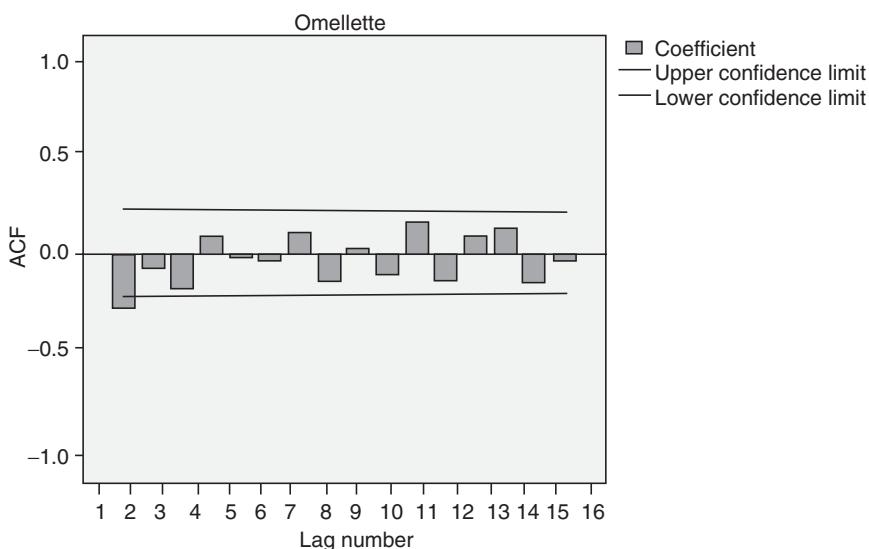
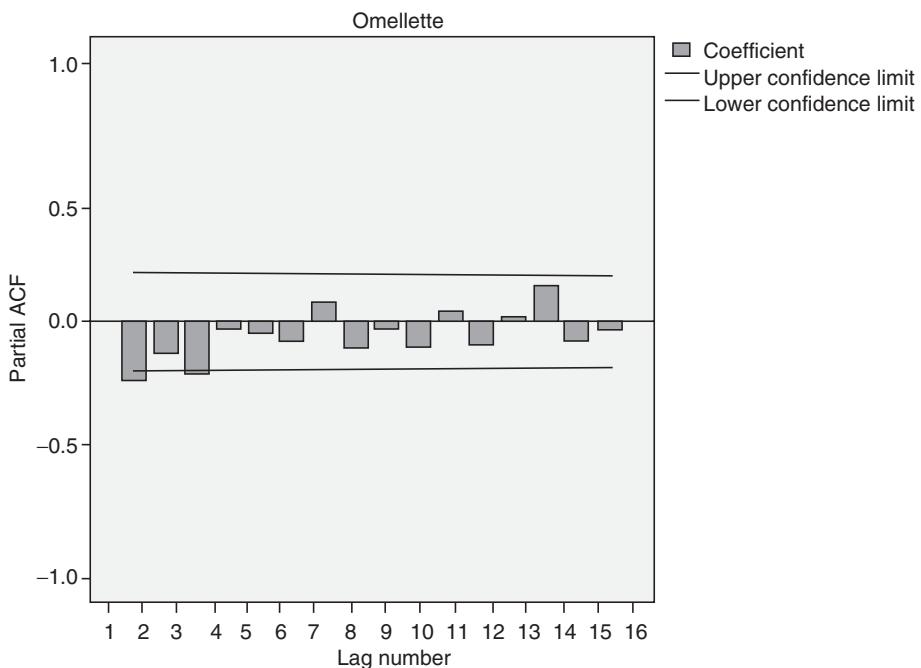


FIGURE 13.17 ACF plot of demand for Omelette after differencing ( $d = 1$ ).



**FIGURE 13.18** PACF plot of demand for Omelette after differencing ( $d = 1$ ).

correlations once it cuts off to zero is not useful and we will ignore them (for example, in PACF plot in Figure 13.17, the partial auto-correlation value with lag 3 is beyond the critical line). However, if the data has seasonal fluctuations, then ACF plot may show consistent spikes as per the periodicity of the seasonal variation. The ARIMA(1, 1, 1) model summary and parameter estimates are shown in Tables 13.29 and 13.30.

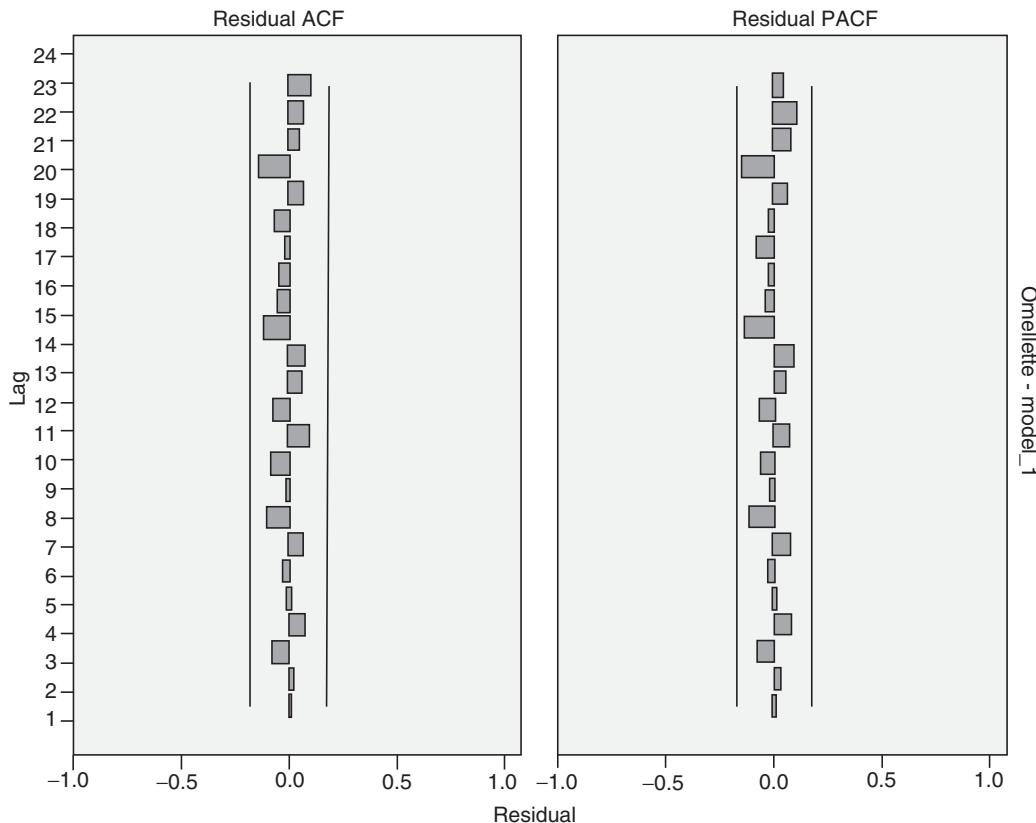
**TABLE 13.29** ARIMA(1, 1, 1) model summary for Omelette demand

Model	Model Fit Statistics			Ljung–Box $Q(18)$		
	R-Squared	RMSE	MAPE	Statistics	Df	Sig.
Omelette-Model_1	0.584	3.439	20.830	10.216	16	0.855

**TABLE 13.30** ARIMA model parameters

		Estimate	SE	T	Sig.
Omelette-Model_1	Constant	0.055	0.137	0.402	0.689
	AR	Lag 1	0.439	0.178	2.475
	Difference		1		0.015
	MA	Lag 1	0.767	0.128	6.004

AR and MA components in Table 13.25 are statistically significant since the corresponding  $p$ -values are less than 0.05. The ACF and PACF of residuals are shown in Figure 13.19 which shows white noise of residuals.



**FIGURE 13.19** ACF and PACF of residuals.

Since the residuals follow white noise, we can use ARIMA(1, 1, 1) model for forecasting.

### 13.14.5 | Ljung–Box Test for Auto-Correlations

Ljung–Box is a test of lack of fit of the forecasting model and checks whether the auto-correlations for the errors are different from zero. The null and alternative hypotheses are given by

- $H_0$ : The model does not show lack of fit
- $H_1$ : The model exhibits lack of fit

The Ljung–Box statistic ( $Q$ -Statistic) is given by (Ljung and Box, 1978)

$$Q(m) = n(n+2) \sum_{k=1}^m \frac{\rho_k^2}{n-k} \quad (13.53)$$

where  $n$  is the number of observations in the time series,  $k$  is the number of lag,  $\rho_k$  is the auto-correlation of lag  $k$ , and  $m$  is the total number of lags.  $Q$ -statistic is an approximate chi-square distribution with  $m - p - q$  degrees of freedom where  $p$  and  $q$  are the AR and MA lags. The  $Q$ -statistic for ARIMA(1, 1, 1) is 10.216 (Table 13.29) and the corresponding  $p$ -value is 0.855 and thus we fail to reject the null hypothesis.  $Q(m)$  measures accumulated auto-correlation up to lag  $m$ .

### 13.15 | POWER OF FORECASTING MODEL: THEIL'S COEFFICIENT

The power of forecasting model is a comparison between Naïve forecasting model and the model developed. In the Naïve forecasting model, the forecasted value for the next period is same as the last period's actual value ( $F_{t+1} = Y_t$ ). Theil's coefficient ( $U$ -statistic) is given by (Theil, 1965)

$$U = \frac{\sum_{t=1}^n (Y_{t+1} - F_{t+1})^2}{\sum_{t=1}^n (Y_{t+1} - Y_t)^2} \quad (13.54)$$

Theil's coefficient is the ratio of the mean squared error of the forecasting model to the MSE of the Naïve model. The value of  $U < 1$  indicates that forecasting model is better than the Naïve forecasting model.  $U > 1$  indicates that the forecasting model is not better than Naïve model. For the data shown in Table 13.14 (demand for avionic system spares), the  $U$ -statistic calculations are shown in Table 13.31.

**TABLE 13.31** U-statistic calculation

Day	$Y_t$	ARMA (1,2) Forecast	$(Y_t - F_t)^2$	Naïve Forecast ( $F_{t+1} = Y_t$ )	$(Y_t - F_t)^2$
31	503	464.8107	1458.423	443	3600
32	688	378.5341	95769.15	503	34225
33	602	444.6372	24763.04	688	7396
34	629	685.8851	3235.909	602	729
35	823	743.5124	6318.281	629	37636
36	671	630.7183	1622.614	823	23104
37	487	649.3491	26357.22	671	33856
		Total	159524.6	Total	140546

The  $U$ -statistic value =  $159524.6 / 140546 = 1.1350$ . That is, ARMA(1, 2) model is not better than Naïve forecasting.

**SUMMARY**

1. Forecasting is one of the important tasks carried out using analytics by many organizations since accurate forecasting is important for taking several decisions such as man-power planning, materials requirement planning, budgeting, and supply chain related issues.
2. Forecasting is carried out on a time-series data in which the dependent variable  $Y_t$  is observed at different time periods  $t$ .
3. Several techniques such as moving average, exponential smoothing, and auto-regressive models are used for forecasting future value of  $Y_t$ .
4. The forecasting models are validated using accuracy measures such as RMSE, MAPE, AIC, and BIC.
5. Simple techniques such as moving average and exponential smoothing may outperform complex regression based models in certain scenarios. Thus, it is important to develop forecasting models using several techniques before selecting the final model.
6. Regression model in the presence of independent variables may outperform other techniques.
7. Auto Regressive (AR) models are regression based models in which dependent variable is  $Y_t$  and the independent variables are  $Y_{t-1}, Y_{t-2}$ , etc.
8. AR models can be used only when the data is stationary.
9. Moving average (MA) models are regression models in which the independent variables are past error values.
10. Auto-regressive integrated moving average (ARIMA) has 3 components: Auto-regressive component with  $p$  lags – AR( $p$ ), moving average component with  $q$  lags – MA( $q$ ), and integration which is differencing the original data to make it stationary.
11. One of the necessary conditions of acceptance of ARIMA model is that the residuals should follow white noise.
12. In ARIMA, the model identification, that is identifying the value of  $p$  in AR and  $q$  in MA, is achieved through auto-correlation function (ACF) and partial auto-correlation function (PACF).
13. The stationarity of time-series data is usually checked using Dickey–Fuller and Augmented Dickey–Fuller test.
14. The overall model accuracy of forecasting model is tested using Ljung–Box test.

**MULTIPLE CHOICE QUESTIONS**

---

1. Seasonality in time-series data is caused due to
  - (a) Changes in macro-economic factors such as recession, unemployment, and so on
  - (b) Festivals and customs in a society
  - (c) Random events that occur over a period of time
  - (d) Changes in customer behaviour driven by new products and promotions
2. In a simple exponential smoothing method, the low value of smoothing constant  $\alpha$  is chosen when
  - (a) The data has high fluctuations around the trend line
  - (b) There is seasonality in the data
  - (c) The data is smooth with low fluctuations
  - (d) There are variations in the data due to cyclical component
3. White noise is
  - (a) Uncorrelated errors with expected value 0.
  - (b) Uncorrelated errors that are constant and do not change with time.
  - (c) Uncorrelated errors that follow normal distribution with mean 0 and constant standard deviation
  - (d) Errors that follow normal distribution with constant mean and standard deviation

4. A stationary process in a time series is a process for which
  - (a) Mean and variance are constant at different time points
  - (b) The time series follows normal distribution with zero mean and constant standard deviation
  - (c) The covariance of the time series depends only on the lag
  - (d) Mean and standard deviation are constant at different time points and the covariance depends only on the lag between the values and is constant for a given lag
5. In a pure auto-regressive process, AR( $p$ ), the value of  $p$  can be identified using
  - (a) Auto-correlation function
  - (b) Partial auto-correlation function
  - (c) Auto-correlation and partial auto-correlation function
  - (d) Ljung–Box test
6. Power of a forecasting model is calculated using
  - (a) Root mean square error (RMSE)
  - (b) Theil's coefficient
  - (c) Mean absolute percentage error (MAPE)
  - (d) Bayesian information criteria (BIC)
7. A necessary condition for accepting a time-series forecasting model is
  - (a) The residuals should follow a normal distribution
  - (b) The residuals should be white noise
  - (c) The residuals should be black noise
  - (d) The residuals should follow a normal distribution and the  $R$ -square should be high
8. In an ARIMA model, differencing is carried out
  - (a) To convert a stationary process to a non-stationary process
  - (b) To convert a non-stationary process to a stationary process
  - (c) To remove seasonal fluctuations from the data
  - (d) To remove cyclical fluctuations from the data
9. Overall fitness of a forecasting model is checked using
  - (a) Durbin–Watson Test
  - (b) Theil coefficient
  - (c) Ljung–Box test
  - (d) Dickey–Fuller test
10. Presence of non-stationarity is checked using
  - (a) Durbin–Watson Test
  - (b) Theil coefficient
  - (c) Ljung–Box test
  - (d) Dickey–Fuller test

## EXERCISES

1. Quarterly demand for certain parts manufactured by Jack and Jill company is shown in Table 13.32.

**TABLE 13.32** Quarterly demand

Year	Quarter	Value
2012	Q1	75
	Q2	60
	Q3	54
	Q4	59
2013	Q1	86
	Q2	65
	Q3	63
	Q4	80
2014	Q1	90

*(Continued)*

**TABLE 13.32** Quarterly demand—Continued

Year	Quarter	Value
2015	Q2	72
	Q3	66
	Q4	85
	Q1	100
2016	Q2	78
	Q3	72
	Q4	93

- (a) Calculate the seasonality index for different quarters using the first 3 years of data.  
 (b) Develop forecasting models using moving average, single exponential smoothing, and an appropriate ARMA model after de-seasonalizing the data (assume multiplicative model,  $Y_t = T_t \times S_t$ ).  
 (c) Forecast the demand for 2015 (all four quarters) using moving average, exponential smoothing, and ARMA. Calculate RMSE, MAPE, and Theil's coefficient.
2. Data on monthly demand for a product over 3 years (between 2013 and 2015) is given in Table 13.33.

**TABLE 13.33** Monthly demand

Month	2013	2014	2015
January	15	23	25
February	16	22	25
March	18	28	35
April	18	27	36
May	23	31	36
June	23	28	30
July	20	22	30
August	28	28	34
September	29	32	38
October	33	37	47
November	33	34	41
December	38	44	53

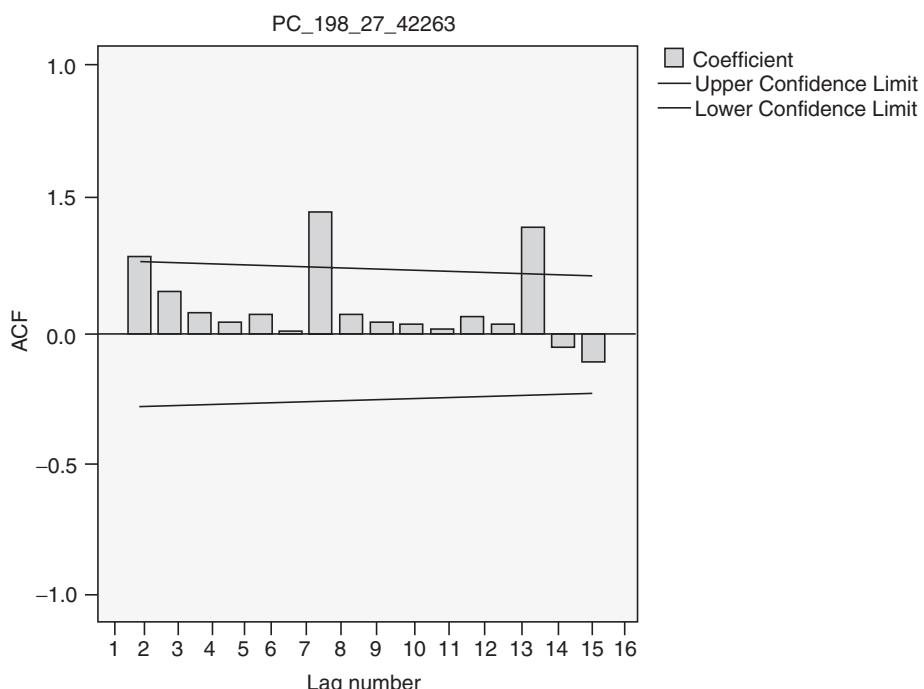
- (a) Calculate the seasonality index using methods of averages.  
 (b) De-seasonalize the data assuming that  $Y_t$  is product of trend and seasonality.  
 (c) Develop the best forecasting model by comparing MAPE of MA, ES, and ARMA models. Compare the models using MAPE and Theil's coefficient.

3. Television rating points of a television program over 30 episodes is shown in Table 13.34.

**TABLE 13.34** Television rating points

Episode	1	2	3	4	5	6	7	8	9	10
TRP	7.98	9.8	9.53	7.23	7.34	9.62	9.8	7.9	8.26	8.17
Episode	11	12	13	14	15	16	17	18	19	20
TRP	8.36	8.5	9.03	9.82	9.77	10.77	9.46	9.31	10.32	9.03
Episode	21	22	23	24	25	26	27	28	29	30
TRP	10.22	10.28	11.99	11.21	9.81	9.35	9.93	11.22	10.4	10.94

- (a) Develop a forecasting model using regression  $Y_t = \beta_0 + \beta_1 t$ , where  $Y_t$  is the TRP at time  $t$ . Is there any trend in the data? Use the regression model developed to answer.
- (b) Is there an auto-correlation in the data? Conduct an appropriate hypothesis test to justify your answer.
- (c) The television channel would like to replace the program with a new program, the average TRP of new program will be 8 points. Based on the model developed, comment whether they should replace the program with a new program.
- (d) Calculate the probability that the TRP for episode 31 will be more than 10.
4. Auto-correlation function and partial auto-correlation functions for a data set are shown in Figures 13.20 and 13.21, respectively.



**FIGURE 13.20** ACF Plot.

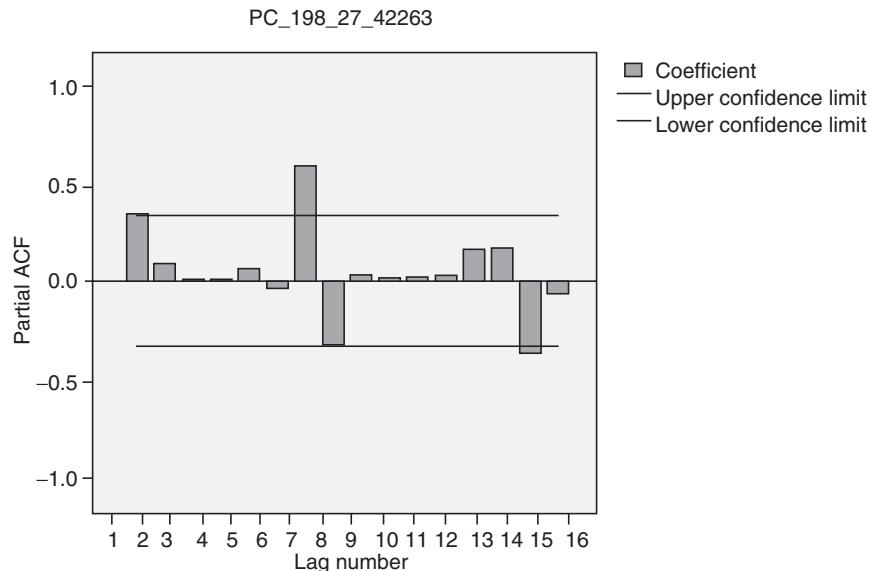


FIGURE 13.21 PACF plot.

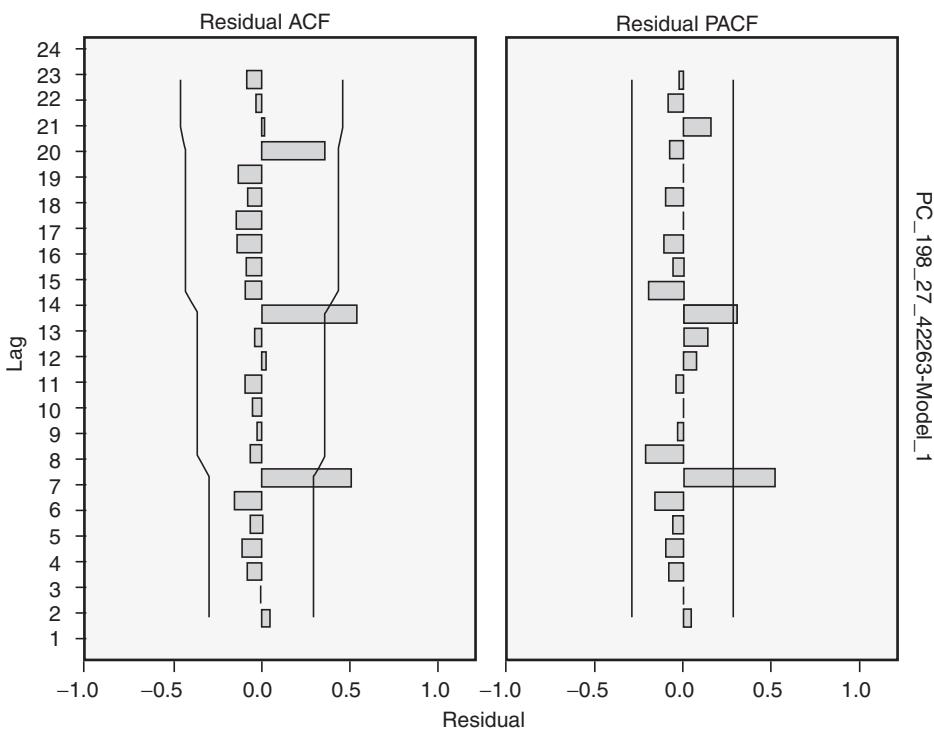


FIGURE 13.22 ACF and PACF of residuals.

- (a) Based on ACF and PACF plot, what values of  $p$  and  $q$  are suitable for auto-regressive and moving average processes?
- (b) The ACF and PACF of residuals are shown in Figure 13.22, what can you conclude from Figure 13.22 about the model?
5. Table 13.35 shows the volume of sales in a retail store on a day and snow fall in inches in the region.

**TABLE 13.35**

Day	1	2	3	4	5	6	7	8	9	10
Sales	72462	607500	816150	973300	744180	255665	1014410	464105	848775	1182225
Snow fall	16.2	20.9	24.84	18.52	5.83	26.74	10.95	22.05	30.35	15.63
Day	11	12	13	14	15	16	17	18	19	20
Sales	646825	656760	597360	559205	159470	225250	1105890	879610	718800	1000875
Snow fall	16.68	15.04	14.11	3.5	5.82	29.18	21.9	18	25.73	18.83
Day	21	22	23	24	25	26	27	28	29	30
Sales	757585	558160	899260	975355	1195165	542225	687160	220905	988275	931930
Snow fall	13.88	23.24	24.77	30.55	12.83	17.64	4.95	26.09	23.46	4.57
Day	31	32	33	34	35	36	37	38	39	40
Sales	218295	387140	289410	448985	1031180	147510	649140	316935	861060	287745
Snow fall	10.08	7.18	11.59	26.88	0.57	17.28	7.53	22.56	6.47	12.77

- (a) Develop a forecasting model using moving average, single exponential smoothing, regression and AR(2). Calculate the MAPE for all 4 models. Which model gives the least MAPE?
- (b) Construct ACF and PACF plot. Develop an ARIMA model [if the model is different from AR(2) model developed in (a)].
- (c) Which model will you recommend to the retail store for forecasting volume of sales?

## Case Study

### Larsen and Toubro – Spare Parts Forecasting<sup>3</sup>

L&T has been in the construction and mining business for the last three decades. It also provides spare parts and after-sales service to customers to improve utilization and helps them to obtain the best possible value. The spare parts business of construction and mining equipment has a profitability of around 30%, while the equipment business has profitability of around 7%. Spare parts availability is very important to the consumer since this affects the

<sup>3</sup> Copyright © Indian Institute of Management, Bangalore. The case was authored by Suhru Kulkarni, Prakash Hegde Ruchi Jaiswal and U Dinesh Kumar, Professor of Quantitative Methods and Information systems prepared this case for classroom discussion. This case is not intended to serve as an endorsement or source of primary data, or to show effective or inefficient handling of decision or business processes. The case was published at the Harvard Business Publishing as part of the IIMB's case collection in 2015. Reproduced with the permission of IIM Bangalore.

**Continued...**

availability of the high-capital cost equipment. However, with over 20,000 different types of spare parts, forecasting the demand of each part is a challenge. Technology changes frequently and superseding parts that have a different material, a different design, or improved quality are developed such that these can replace the previous parts. New machines that are developed require additional spare parts. Further, demand is quite seasonal since most of the customers plan their annual overhaul during the monsoon season, when construction is virtually at a standstill.

— Vijaya Kumar, DGM, L&T Construction  
and Mining Business (April 2014)

Monday, April 21, 2014: Vijaya Kumar walked out of his cabin trying to clear his head that was filled with numbers. He was going over the demand trends and forecasts for around 20,000 spare parts of the various construction and mining equipment sold by Larsen and Toubro (L&T). Vijaya Kumar was the Deputy General Manager of the Supply Chain Department of L&T's Construction and Mining Business (CMB). L&T has been India's largest technology, engineering, construction, and manufacturing company. CMB provided heavy construction and mining equipment to its customers, along with support services and spare parts. The supply of spare parts was critical since the customer would face severe losses in the event of equipment unavailability. Forecasting was done *ad hoc*, based on the experience of the planning personnel. The value of each spare part ranged from INR 10 to INR 8 million (USD 1 ≈ INR 60 in April 2014). Maintaining the balance of the spare parts inventories was critical since unavailability would result in loss of revenues, decreased profitability, and increased customer dissatisfaction, and would also give rise to the spurious products industry. Excess inventory would lead to high inventory carrying costs, working capital lock-in, and the possibility of the spare parts becoming obsolete. Kumar had to arrive at an accurate forecasting methodology for the 20,000-odd spare parts that CMB had to supply. In theory, 20,000 spare parts called for 20,000 forecasting models; however, such a large number of models would be very time-consuming as well as expensive to develop and manage. Kumar wanted to build the forecasting model quickly so that he could roll out the forecasting strategy on a pan-India basis within a few weeks.

### **Construction and Mining Equipment Industry In India**

Construction and mining equipment mainly included earthmoving equipment and material-handling equipment, along with a variety of other machineries (**Exhibit 1**). Prior to 1960, India imported all of its construction and mining equipment owing to the lack of indigenous manufacturing facilities. In 1964, Bharat Earth Movers Ltd. (BEML), a public sector enterprise, was established to produce dozers and dumpers under technology license from LeTorneu Westinghouse (USA) and Komatsu (Japan). Another private sector enterprise, Hindustan Motors, forayed into this sector in 1969. Hindustan Motors had technological collaboration with Terex, UK. Several foreign players such as Case, Caterpillar, Inggersoll Rand, Komatsu, JCB, Hitachi, Volvo, and Lieber entered the country after the 1991 economic reforms, either through joint ventures with Indian companies or by setting up wholly owned subsidiaries.

## Continued...

**Case Study**

The industry size was estimated to be around INR 153 billion (USD 2.55 billion); it had grown at a CAGR of around 10% during the fiscal period of 2009–2013 owing to a high growth of 33% during FY 2009–2010. The annual growth rate was around 3.4% during 2010–2013.<sup>4</sup> The growth of the construction and mining equipment industry depended on the construction and mining activities in the country, which had considerably slowed down owing to the economic slowdown, high interest rates, and the slowdown in policy making by the Indian government. The construction and mining sectors reported growth of around 5.9% and 0.4%, respectively, in FY 2012–2013. The construction sector comprised residential and commercial real estate along with infrastructure construction. The demand for residential and commercial estate had decreased. There was a slowdown in the infrastructure sector owing to delays in obtaining government clearances, trouble with land acquisition, and the high cost of capital. The mining sector largely included coal, iron ore, and limestone mining; this sector was affected by regulatory uncertainties. The proposed infrastructure spending of USD 1 trillion under the 12<sup>th</sup> Five Year Plan (2012–2017) and regulatory clarity over mining licenses were expected to revive the construction and mining industry; the projected growth in the infrastructure sector was estimated to be not more than 5% per annum. Similar growth was expected in the construction and mining equipment industry.

The construction and mining equipment industry mainly involved large players owing to the need for high upfront capital investments and technical expertise. BEML, L&T, Caterpillar, JCB India, Komatsu, Case, Volvo, Telcon, Escorts, Tata-Hitachi, Kobelco, Liebherr, Ingersoll Rand, and Voltas were some of the major players in the industry in India.

In the construction industry, the customers were mainly unorganized players, since construction activities were sub-contracted to smaller players by the large construction companies. However, in the mining industry, the customers were large coal mining companies (such as Coal India Ltd. and its subsidiaries), large steel manufacturers (for iron ore) (such as Steel Authority of India, Tata Steel, and Jindal Steel), and cement manufacturers (such as ACC, Ambuja Cements, etc.). Construction and mining equipment were expensive, and the life of each piece of equipment was around 20,000 hours (around 3.5 years, working in two shifts). The customers required high equipment availability to enable them to recover their capital costs. The maintenance and availability of spare parts were critical for ensuring equipment availability. Further, genuine spare parts from the original equipment manufacturer (OEM) had to be used to avoid damage to the equipment. Some customers purchased spare parts from local vendors for one of the two reasons – unavailability of spare parts with the OEM, or the lower cost of the local/spurious spare parts, which affected the performance of the equipment. Thus, service and spare parts were critical elements of the construction and mining equipment industry.

### L&T: The Indian Engineering Giant

L&T was founded in Bombay (Mumbai) in 1938 by two Danish engineers, Henning Holck-Larsen and Soren Kristian Toubro. They started their operations by representing Danish dairy

---

<sup>4</sup> Source: India's Construction and Mining Equipment Industry, D&B, August 2013.

**Continued...**

equipment manufacturers. World War II stopped Danish supplies, which forced the founders to manufacture equipment indigenously. The war created opportunities for repairing ships, which led to the formation of repair and fabrication shops. During 1944–1946, L&T entered into several foreign collaborations. After India gained independence in 1947, there was a growing demand for equipment across industries. L&T expanded and set up offices across the eastern (Calcutta), southern (Madras), and northern (New Delhi) regions of India. In 1950, L&T became a public company, and it grew rapidly in the 1960s. Toubro and Larsen retired in 1962 and 1978, respectively. The company grew into an engineering major under the guidance of several eminent leaders.

In 2014, L&T established presence in several businesses – infrastructure, defense, aerospace, hydrocarbons, heavy engineering, construction, power, mining and metallurgy, electrical and automation products, machinery and industrial products, information technology, financial services, ship building, and railway projects. L&T had design-to-build capacities for most of its businesses. It had 137 subsidiaries and 16 associated companies; it had not only a pan-India presence but also a global presence across 30 countries. The L&T Group had gross revenues of INR 752 billion (USD 13 billion) in FY 2013.<sup>5</sup>

### L&T's Construction and Mining Business

The Construction and Mining Business (CMB) formed a part of the Machinery and Industrial Products (MIP) business at L&T (**Exhibit 2**). The MIP segment earned revenues of INR 23 billion, which formed around 4% of L&T's total revenues. However, the MIP segment had gross profit margins of around 16.3% as against L&T's overall gross profit margin of 13.2%.<sup>6</sup> Thus, even though MIP was a small segment in terms of revenues, it was important from the profitability perspective. CMB was formed in 1998 as a 50:50 joint venture between L&T and Komatsu Asia & Pacific, Singapore, which was a wholly owned subsidiary of Komatsu, Japan. In April 2013, L&T bought out Komatsu's 50% stake.<sup>7</sup>

L&T's CMB sold equipment such as dozer shovels, dozers, dumpers, hydraulic excavators, motor graders, pipe layers, surface miners, tipper trucks, wheel dozers, and wheel loaders (**Exhibit 3**). CMB also provided equipment installation and commissioning services as well as other maintenance services. CMB did not manufacture the equipment that it sold and serviced; it had tie-ups with OEMs such as Komatsu, Scania, and so on for sourcing the equipment. Additionally, CMB sold and serviced construction and mining equipment that was manufactured in other L&T divisions.

CMB involved different verticals such as sales and marketing, services, and supply chain. The sales and marketing vertical was responsible for marketing and completing the sale orders that were

<sup>5</sup> Source: <http://www.larsentoubro.com/>.

<sup>6</sup> Source: L&T Annual Report, 2012–13.

<sup>7</sup> Source: [http://articles.economictimes.indiatimes.com/2013-04-13/news/38511372\\_1\\_larsen-toubro-construction-equipment-brand-equity](http://articles.economictimes.indiatimes.com/2013-04-13/news/38511372_1_larsen-toubro-construction-equipment-brand-equity)

**Continued...**

placed with the OEMs. The OEM invoiced the customer and dispatched the equipment. Subsequently, CMB's service team would install and commission the equipment for the customer. The service team also took care of servicing during the warranty period and dealt with any other kind of servicing-related assistance that the customer required. The supply chain team was responsible for the availability of spare parts and the related internal logistics.

Case Study

### Spare Parts Supply Chain at CMB

CMB supplied parts across the country through its central warehouse at Nagpur. The CMB group had four fully equipped service stations at Chennai, Delhi, Durgapur, and Pune, which catered to the southern, northern, eastern, and western regions of the country, respectively. The facility at Nagpur catered to the central region. There were 58 sub-service centres under the fully equipped service centres. In addition to this, CMB had 26 dealers who had 84 outlets across the country. The spare parts of the construction equipment were generally supplied through the dealers, while the mining equipment spare parts were supplied via a mix of service centres and dealer outlets.

The spare parts were classified into six main categories: filters, engine maintenance parts, seals and hoses, v-belts, undercarriage, and ground engaging tools (**Exhibit 4**). Each category included different types of spare parts, and each type consisted of several varieties with different technical specifications. On average, a major piece of equipment would consist of around 1,100 unique parts. Thus, there were more than 20,000 spare parts across the different types of equipment sold by L&T.

The demand for the spare parts would vary owing to a variety of reasons. Some spare parts would be required because of operational wear and tear, while others would be needed owing to breakdown. Customers followed different maintenance strategies such as preventive or breakdown maintenance, which caused variation in demand. CMB provided warranties for the equipment; in the instance of any breakdown during the warranty period, spare parts would be required. Warranties were generally provided for 3,000 hours or 1 year of equipment operations; the typical life of each piece of equipment was around 20,000 hours. Further, several customers opted for extended warranty, which added to the variation in demand. During the extended warranty period, L&T had to support the machine similar to what was done during the warranty period. Annual Maintenance Contracts (AMC) guaranteed the availability of equipment to customers. Availability of spare parts played a critical role in fulfilling AMC services. Additionally, the sales and marketing team sometimes offered spare parts free of cost. The sales team would offer some free spare parts to the customers as per the prevailing offers made by the competition. Some of the customers would not opt for these offers and would demand equivalent value. In such instances, L&T would offer them equivalent value for the future purchase of any other spare parts from L&T (for this value). This was similar to providing coupons that were equal in value to the free spare parts offered by L&T. These coupons could be redeemed for spare parts purchased from L&T. The customer benefitted from the option of using the value gained from the 'free of cost' offer for future purchases. However, the purchases made by the customers using such coupons were generally one-time purchases, and the probability of the repetition of this demand was low. Thus, the variability in demand increased with such 'free of cost' offers.

**Continued...**

The after-sales business in the machine and plant construction industry generally accounted for approximately 25% of the total sales (with two-thirds being derived from the sales of spare parts and one-third from services), while it accounted for almost 50% of the total profits.<sup>8</sup> At CMB, spare parts contributed 25% of the total revenues; the rest was contributed by equipment sales. The profitability of the spare parts business was around 30%, while the profitability of the equipment sales business was around 7%. The unavailability of spare parts led to not only the loss of a transaction but also the potential loss of a customer, since the customer preferred immediate replacement of the damaged spare parts to avoid losses to his/her business.

### Forecasting Demand for Spare Parts

Kumar was looking at the data for the period April 2009–April 2013. He had monthly details of the demand for each of the spare parts. It was difficult and very expensive to forecast the individual demand for 20,000 spare parts. Therefore, the CMB team had categorized the spare parts in several ways. The first categorization was based on the frequency of demand, according to which three categories – fast (F), medium (M), and slow (S) – were formed. This was known as the FMS categorization. The second categorization was based on the value of the product. Three categories were formed according to value: high (H), medium (M), and low (L); this was known as the HML categorization. The third categorization was based on the well-known ABC analysis according to which the spare parts belonging to category A constituted 70% of the sales, the spare parts belonging to category B constituted 25% of the sales, while category C spare parts constituted 5% of the sales. The spare parts were further classified based on combinations of these three categories and were ranked accordingly. For example, a category A spare part that was fast-moving (F) and had high value (H) was assigned the top rank; this combination was termed ‘AHF’ (**Exhibit 5**). According to Kumar:

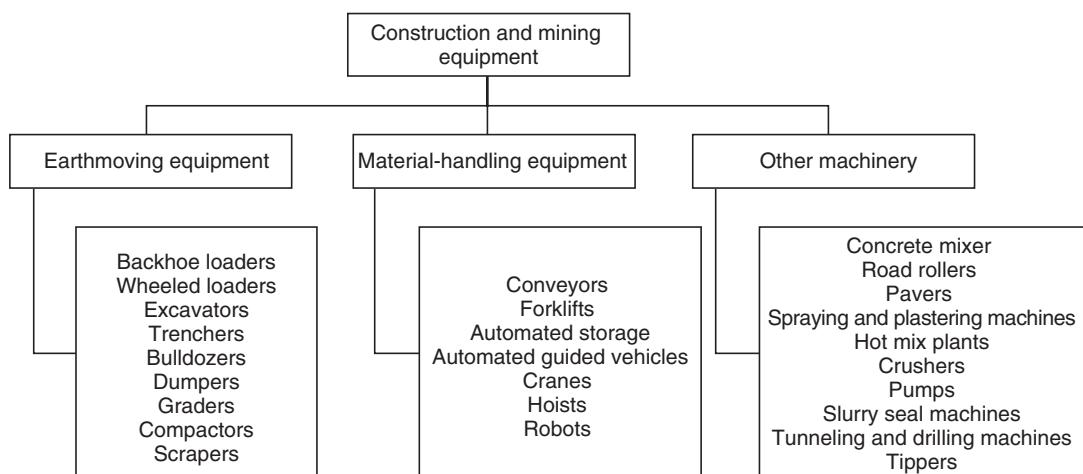
We used several analytical tools such as exponential smoothing, autoregressive integrated moving average (ARIMA) model, and Croston’s method for forecasting the different categories of spare parts. We figured that one forecasting model would not be suitable for all the spare parts since the pattern of demand varied across the spare parts.

Kumar’s team wanted to develop forecasting models for spare parts that would result in less than 10% error. However, Kumar wondered whether it was possible to develop such a model for all the spare parts and whether this model would work in the face of constantly changing industry trends. Was there any additional data that could be incorporated into the model to make it more robust? What strategies should be followed by the CMB team to ensure better availability and reduced inventory costs?

<sup>8</sup> Stephen M. Wagner, Ruben Jonke, and Andreas B. Eisingerich, A Strategic Framework for Spare Parts Logistics, *California Management Review*, 2012, 54(4), 69.

Continued...

Case Study



**EXHIBIT 1** Classification of construction and mining equipment. Source: India's construction and mining equipment Industry, D&B, August 2013.

Case Study •

Continued...

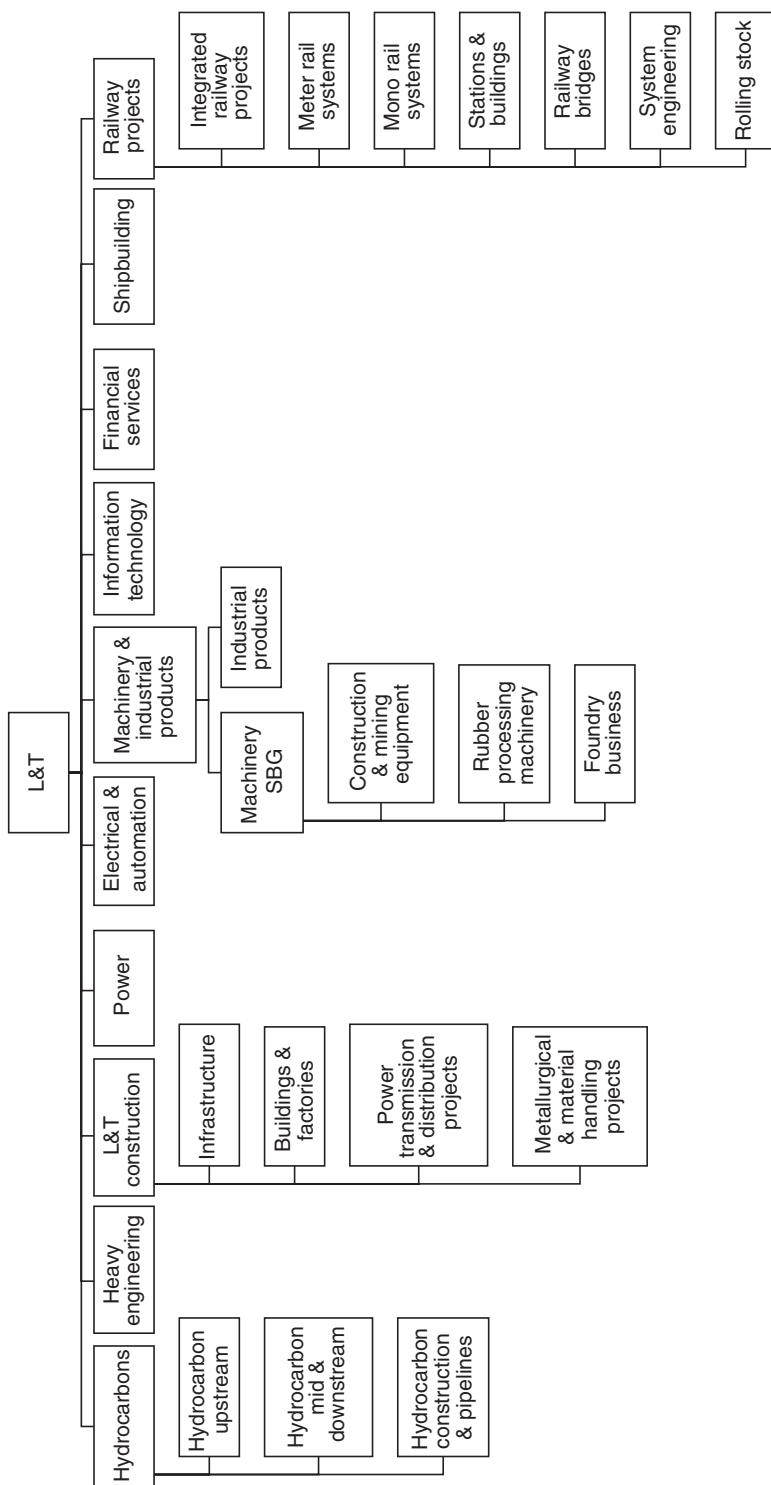


EXHIBIT 2 L&T's organizational structure. Source: L&T.

Continued...

Case Study



Dozers



Dumpers



Hydraulic Excavators



Motor Graders



Surface Miners



Tipper Truck

**EXHIBIT 3** L&T CMB's equipment. Source: L&T.

Continued...

**EXHIBIT 4** L&T CMB's spare parts

Filters	Engine Maintenance Parts	Seals & Hoses	Undercarriage	Ground Engaging Tools	V-Belts
Engine Filters	Gasket Kits	Hydraulic Hoses	Track-Links	Bucket Teeth	
1000-Hour Filters	Rain Caps	O-Rings	Sprocket Teeth	Cutting Edges	
Fuel Filters	Exhaust Pipe	Oil Seals	Track Shoes	Blades	
Oil Filters	Mufflers	Fuel Hoses	Rollers	Side Cutters	
Air Cleaner	Cylinder Liners	Dust Seals	Idlers		
Corrosion Resistor	Pistons	Seal Washers			
Hydraulic Filter	Piston Rings	Low Pressure Heads			
	Thermostat	Water Hoses			
	Fuel Water Separators	Hose Clamps			
		Back-up Rings			

Source: L&T CMB.



Fuel Filters



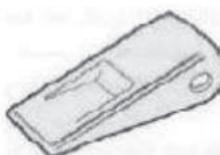
Gasket Kit



O-Rings



Rollers



Bucket Tooth

Continued...

**EXHIBIT 5** Categorization of spare parts

FMS Categorization (Demand)	
Fast (F)	Demand > 8 per month
Medium (M)	8 per month ≥ Demand > 4 per month
Slow (S)	Demand ≤ 4 per month

ABC Categorization (Sales)	
A	70% of sales
B	25% of sales
C	15% of sales

HML Categorization (Value)	
High (H)	Value ≥ INR 50,000
Medium (M)	INR 50,000 > Value > INR 10,000
Low (L)	Value ≤ INR 10,000

ABC	HML	FMS	Comb	Rank
A	H	F	AHF	1
A	M	F	AMF	2
A	L	F	ALF	3
B	H	F	BHF	4
B	M	F	BMF	5
B	L	F	BLF	6
C	H	F	CHF	7
C	M	F	CMF	8
C	L	F	CLF	9
A	L	M	ALM	10
B	M	M	BMM	11
B	L	M	BLM	12
C	M	M	CMM	13
C	L	M	CLM	14
A	H	M	AHM	15
A	M	M	AMM	16
B	H	M	BHM	17
C	H	M	CHM	18
A	H	S	AHS	19
A	M	S	AMS	20

ABC	HML	FMS	Comb	Rank
A	L	S	ALS	21
B	H	S	BHS	22
B	M	S	BMS	23
B	L	S	BLS	24
C	H	S	CHS	25
C	M	S	CMS	26
C	L	S	CLS	27

Source: L&T construction and mining business

Continued...

**CASE QUESTIONS (USE THE DATA PROVIDED)**

1. What strategy should Vijaya Kumar adopt for developing forecasting model for demand estimation of 20,000 spare parts?
2. Develop forecasting models for data provided in the Excel sheet titled “L&T Spare Parts Forecasting” and discuss the choice for using a particular forecasting model.
3. Which forecasting techniques should L&T use to forecast different spare items?

**REFERENCES**

1. Ali F (2017), “Amazon could Drive 80% of U.S. e-Commerce Growth”, *Digital 360 Commerce*, March 8 2017. Available at <https://www.digitalcommerce360.com/2017/03/08/amazon-drive-80-u-s-e-commerce-growth-next-year/>, accessed on 10 May 2017.
2. Box G E P and Jenkins G M (1970), “*Time Series Analysis, Forecasting and Control*”, Holden Day, San Francisco.
3. Chatfield C (1986), “Simple is Best?”, Editorial in the *International Journal of Forecasting*, **2**, 401–402.
4. Croston J D (1972), “Forecasting and Stock Control for Intermittent Demands”, **23**(3), 289–303.
5. Dickey D A and Fuller W A (1979), “Distribution of Estimation of Auto-Regressive Time Series with a Unit Root”, *Journal of the American Statistical Association*, **74**, 427–431.
6. Fuller W. (1979), “Introduction to Statistical Time Series”, John Wiley and Sons, New York.
7. Hill K (2011), “Extreme Engineering: The Boeing 747”, *Science Based Life – Add Little Reason to Your Day*, 25 July 2011, available at <https://sciencebasedlife.wordpress.com/2011/07/25/extreme-engineering-the-boeing-747/> accessed on 10 May 2017.
8. Ljung G M and Box G E P (1978), “On a measure of Lack of fit in Time Series Models”, *Biometrika*, **65**(2), 297–303.
9. Makridakis S, Wheelwright S C, and Hyndman R J (1998), “*Forecasting – Methods and Applications, Third Edition*”, John Wiley & Sons, USA.
10. Parker G C and Segura E L (1971), “How to get a Better Forecast”, *Harvard Business Review*, March-April 1971, 99–109.
11. Taylor, J W (2011), “Multi-Item Sales Forecasting with Total and Split Exponential Smoothing”, *The Journal of the Operational Research Society*, **62**(3), 555–563.
12. Theil H (1965), “*Economic Forecasts and Policy*”, North Holland, Amsterdam.
13. Yaffee R A and McGee M (2000), “*An Introduction to Time Series Analysis and Forecasting: With Applications of SAS and SPSS*”, Academic Press, New York.
14. Winters P R (1960), “Forecasting Sales by Exponentially Weighted Moving Averages”, *Management Science*, **6**(3), 324–342.

# 14

# CLUSTERING

“Barn’s burnt down now I can see the moon”

— Mizuta Masahide

## LEARNING OBJECTIVES

- LO 14-1** Understand the role of clustering and its importance in analytics.
- LO 14-2** Learn different types of clustering techniques.
- LO 14-3** Understand various distance measures such as Euclidean distance, Minkowski distance, Cosine similarity, Jaccard distance, Gower's similarity and its applications in clustering.
- LO 14-4** Identify cluster characteristics and its importance in designing strategies.
- LO 14-5** Understand how clustering helps data scientists with customer segmentation and personalized actions.

## ESSENCE OF CLUSTERING

Clustering is one of the most frequently used analytics applications. Clustering helps data scientists to create homogeneous group of customers/entities for better management of customers. In many analytics projects, once the data preparation is complete, clustering is usually carried out before applying other analytical models. Clustering is a divide-and-conquer strategy which divides the data set into homogenous groups which can be further used to prescribe right strategy for different groups. In clustering, the objective is to ensure that the variation within a cluster is minimized whereas the variation between clusters is maximized.



*Clustering is usually one of the first tasks performed in most analytics projects. It helps data scientists to analyze individual clusters further.*

## 14.1 | INTRODUCTION TO CLUSTERING

Clustering is an important task in analytics in which the data (customers or entities) is grouped into finite subsets such that each subset is a homogeneous group of entities. Many analytics projects may start first with clustering after performing descriptive statistics and visualization on the data, since it assists data scientists to apply appropriate strategies for different clusters identified through cluster characteristics.

Clustering algorithms attempt to solve a classification problem, in which the objective is to find different classes that exist in the data. The main difference between clustering algorithms and other classification techniques such as logistic regression and classification trees is that clustering algorithms are unsupervised learning algorithms (classes are not known *a priori*) whereas logistic regression and classification tree are supervised learning algorithms (where classes are known *a priori* in the training data). Another important difference between clustering and classification is that clustering is descriptive analytics whereas classification is usually a predictive analytics algorithm. The end objective of clustering is to create heterogeneous subsets (clusters) from the original data set such that each subset is homogeneous within the cluster and identify the characteristics that differentiate the subsets. Clusters can be classified into the following four categories:

1. **Non-overlapping clusters:** Cluster in which each observation belongs to only one cluster. Non-overlapping clusters are more frequently used clustering techniques in practice.
2. **Overlapping clusters:** An observation may belong to more than one cluster.
3. **Probabilistic clusters:** An observation may belong to a cluster according to a probability distribution.
4. **Hierarchical clustering:** Hierarchical clustering creates subsets of data similar to a tree-like structure in which the root node corresponds to the complete set of data. Branches are created from the root node to split the data into heterogeneous subsets (clusters).

Clustering algorithms use different distance or dissimilarity measures to derive different clusters. The type of distance/dissimilarity measure used plays a crucial role in the final cluster formation. Higher distance would imply that observations are dissimilar, whereas higher similarity would indicate that the observations are similar.

## 14.2 | DISTANCE AND DISSIMILARITY MEASURES USED IN CLUSTERING

Clustering techniques assume that there are subsets in the data that are dissimilar. One approach for measuring dissimilarity is through distances measured using different metrics.

### 14.2.1 | Euclidean Distance

Euclidean is one of the frequently used distance measures when the variable is either in interval or ratio scale. Assume that the data has  $n$  attributes ( $n$  variables or  $n$  features). Then the Euclidean distance between two  $n$ -dimensional observations  $X_1(x_{11}, x_{12}, \dots, x_{1n})$  and  $X_2(x_{21}, x_{22}, \dots, x_{2n})$  is given by

$$D(X_1, X_2) = \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + \dots + (x_{1n} - x_{2n})^2} \quad (14.1)$$

Euclidean distance is a straight line distance between two points  $X_1$  and  $X_2$ . Smaller Euclidean distance implies that observations  $X_1$  and  $X_2$  are similar and are likely to be a part of the same cluster, whereas large Euclidean distance implies high dissimilarity between  $X_1$  and  $X_2$  and thus they are likely to be a part of different clusters. Table 14.1 has information about 20 wines sold in the market along with their alcohol and alkalinity of ash content.

**TABLE 14.1** Alcohol and alkalinity of ash in wine

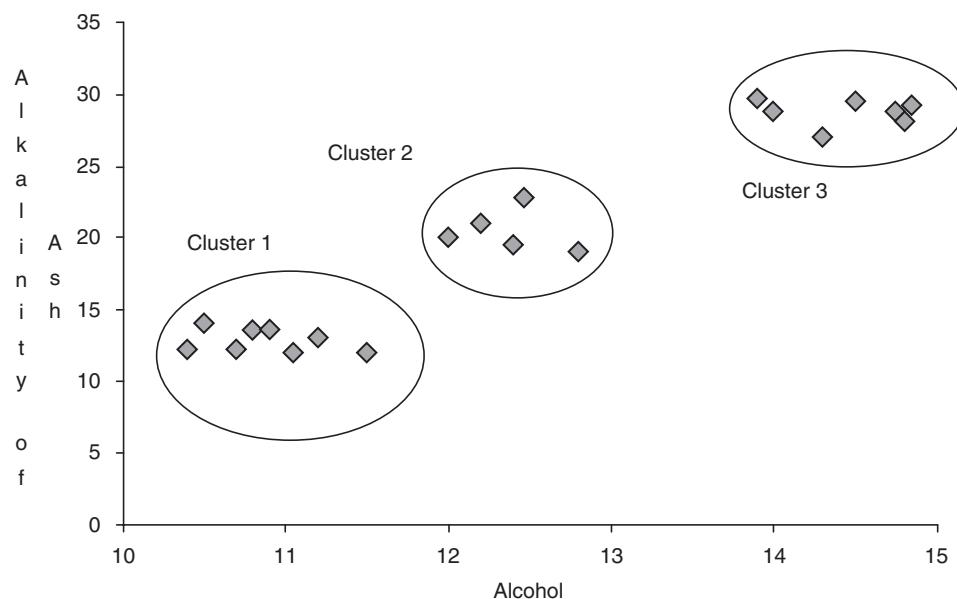
Wine	Alcohol	Alkalinity of Ash	Wine	Alcohol	Alkalinity of Ash
1	14.8	28	11	10.7	12.2
2	11.05	12	12	14.3	27
3	12.2	21	13	12.4	19.5
4	12	20	14	14.85	29.2
5	14.5	29.5	15	10.9	13.6
6	11.2	13	16	13.9	29.7
7	11.5	12	17	10.4	12.2
8	12.8	19	18	10.8	13.6
9	14.75	28.8	19	14	28.8
10	10.5	14	20	12.47	22.8

The plot of 20 wines with respect to alcohol and alkalinity of ash is shown in Figure 14.1. It is very clear from Figure 14.1 that there are three clusters and the wines in these clusters are given by

$$\text{Cluster 1} = \{2, 6, 7, 10, 11, 15, 17, 18\}$$

$$\text{Cluster 2} = \{3, 4, 8, 13, 20\}$$

$$\text{Cluster 3} = \{1, 5, 9, 12, 14, 16, 19\}$$

**FIGURE 14.1** Clusters of wine based on alcohol and ash content.

The Euclidean distance between any two observations within the cluster will be lesser than the observations between clusters. For example, consider the distance between wines 2 and 6 which are part of cluster 1. The Euclidean distance between wine 2 and wine 6 is given by

$$D(\text{wine 2}, \text{wine 6}) = \sqrt{(11.05 - 11.20)^2 + (12 - 13)^2} = 1.011$$

Euclidean distance between 1 (part of cluster 3) and 2 (part of cluster 1) is given by

$$D(\text{wine 1}, \text{wine 2}) = \sqrt{(14.80 - 11.05)^2 + (28 - 12)^2} = 16.433$$

That is, the Euclidean distance between observations within the cluster is smaller than between the clusters as evident from distance between wines 1, 2 and wines 2, 6. The centroid (average of different parameter values) of clusters is given by

$$\text{Centroid of cluster 1} = \left( \frac{\begin{array}{c} 11.05 + 11.2 + 11.5 + 10.5 + 10.7 + 10.9 + 10.4 + 10.8 \\ 8 \end{array}}{\frac{12 + 13 + 12 + 14 + 12.2 + 13.6 + 12.2 + 13.6}{8}}, \right) = (10.88, 12.83) \quad (14.2)$$

$$\text{Centroid of cluster 2} = \left( \frac{\begin{array}{c} 12.2 + 12 + 12.8 + 12.4 + 12.47 \\ 5 \end{array}}{\frac{21 + 20 + 19 + 19.5 + 22.8}{5}}, \right) = (12.37, 20.46) \quad (14.3)$$

$$\text{Centroid of cluster 3} = \left( \frac{\begin{array}{c} 14.80 + 14.5 + 14.75 + 14.3 + 14.85 + 13.9 + 14 \\ 7 \end{array}}{\frac{28 + 29.5 + 28.8 + 27 + 29.2 + 29.7 + 28.8}{7}}, \right) = (14.44, 28.71) \quad (14.4)$$

Distances between different centroids are given in Table 14.2.

**TABLE 14.2** Distances between cluster centroids

Euclidean Distance	Centroid of Cluster 1	Centroid of Cluster 2	Centroid of Cluster 3
Centroid of cluster 1	0	7.774	16.278
Centroid of cluster 2	7.774	0	8.506
Centroid of cluster 3	16.278	8.506	0

A new data can be added to the three clusters identified by calculating the distance between the centroid of the cluster and the data. For example, consider a new wine (say wine 21) for which the alcohol and ash content are 12 each. Then the distance between centroids of various clusters and wine 21 are

$$D(\text{cluster 1}, \text{wine 21}) = \sqrt{(10.88 - 12)^2 + (12.83 - 12)^2} = 1.38$$

$$D(\text{cluster 2}, \text{wine 21}) = \sqrt{(12.37 - 12)^2 + (20.46 - 12)^2} = 8.47$$

$$D(\text{cluster 3}, \text{wine 21}) = \sqrt{(14.44 - 12)^2 + (28.71 - 12)^2} = 16.89$$

Since the distance between cluster 1 and wine 21 is the smallest among the three clusters, wine 21 will be classified under cluster 1. A primary disadvantage of Euclidean distance is that it can be used only when the attributes are numerical measures (ratio and interval scale).

#### 14.2.2 | Standardized Euclidean Distance

One of major issues with Euclidean distance is that the scale of different attributes within the data can be different. Let  $X_{1k}$  and  $X_{2k}$  be two attributes of the data (where  $k$  stands for the  $k^{\text{th}}$  observation in the data set). It is possible that the range of  $X_{1k}$  can be much larger compared to  $X_{2k}$ , resulting in skewed Euclidean distance value (for example  $X_1$  is salary and  $X_2$  is family size). An easier way of handling the potential bias is to standardize the data using the following equation:

$$\text{Standardized value of the attribute} = \left( \frac{X_{ik} - \bar{X}_i}{\sigma_{X_i}} \right) \quad (14.5)$$

where  $\bar{X}_i$  and  $\sigma_{X_i}$  are, respectively, the mean and standard deviation of  $i^{\text{th}}$  attribute.

#### 14.2.3 | Manhattan Distance (City Block Distance)

Euclidean distance may not be appropriate while measuring distance between different locations (for example, distance between two shops in a city). In such cases, we use Manhattan distance, which is given by

$$DM(X_1, X_2) = \sum_{i=1}^n |X_{1i} - X_{2i}| \quad (14.6)$$

#### 14.2.4 | Minkowski Distance

Minkowski distance is the generalized distance measure between two cases in the data set and is given by

$$\text{Minkowski } D(X_1, X_2) = \left( \sum_{i=1}^n |X_{1i} - X_{2i}|^p \right)^{1/p} \quad (14.7)$$

When  $p = 1$ , Minkowski distance is same as the Manhattan distance and for  $p = 2$ , Minkowski distance is same as the Euclidean distance.

#### 14.2.5 | Jaccard Similarity Coefficient (Jaccard Index)

Jaccard similarity coefficient (JSC) or Jaccard index (Real and Vargas, 1996) is a measure used when the data is qualitative, especially when attributes can be represented in binary form. JSC is one of the popular measures in collaborative filtering techniques and is frequently used in developing recommender systems. JSC for two  $n$ -dimensional data ( $n$  attributes),  $X_1$  and  $X_2$ , is given by

$$\text{Jaccard}(X_1, X_2) = \frac{n(X_1 \cap X_2)}{n(X_1 \cup X_2)} \quad (14.8)$$

where  $n(X_1 \cap X_2)$  is the number of attributes that belong to both  $X_1$  and  $X_2$  (that is,  $X_1 \cap X_2$ ),  $n(X_1 \cup X_2)$  is the number of attributes that belong to either  $X_1$  or  $X_2$  (that is,  $X_1 \cup X_2$ ). Prior to applying JSC, the data has to be pre-processed to convert each attribute in the data into a binary or Boolean representation. For example, consider movie DVD purchases made by two customers as given by the following sets:

Customer 1 = {Jungle Book (JB), Iron Man (IM), Kung Fu Panda (KFP), Before Sunrise (BS), Bridge of spies (BoS), Forrest Gump (FG)}

Customer 2 = {Casablanca (C), Jungle Book (JB), Forrest Gump, Iron Man (IM), Kung Fu Panda (KFP), Schindler's List (SL), The God Father (TGF)}

In this case, each movie is an attribute. The purchases made by the two customers are shown in Table 14.3. Where 1 implies purchase of DVD and 0 otherwise.

**TABLE 14.3** Binary representation of customer purchases

Movie Title	BS	BoS	C	FG	IM	JB	KFP	SL	TGF
Customer 1	1	1	0	1	1	1	1	0	0
Customer 2	0	0	1	1	1	1	1	1	1

The JSC is given by

$$\text{JSC} = \frac{n(\text{customer 1} \cap \text{customer 2})}{n(\text{customer 1} \cup \text{customer 2})} = \frac{4}{9} = 0.44$$

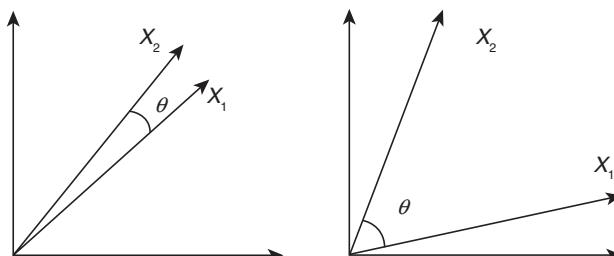
Higher the Jaccard coefficient, higher the similarity between two observations being compared. The value of JSC lies between 0 and 1.

#### 14.2.6 | Cosine Similarity

Consider two cases  $X_1$  and  $X_2$  with  $n$ -attributes  $X_1 = (x_{11}, x_{12}, \dots, x_{1n})$  and  $X_2 = (x_{21}, x_{22}, \dots, x_{2n})$ . Then the cosine similarity between  $X_1$  and  $X_2$  is given by

$$\text{Similarity } (X_1, X_2) = \cos(\theta) = \frac{X_1 \cdot X_2}{\|X_1\| \cdot \|X_2\|} = \frac{\sum_{i=1}^n X_{1i} \times X_{2i}}{\sqrt{\sum_{i=1}^n X_{1i}^2} \times \sqrt{\sum_{i=1}^n X_{2i}^2}} \quad (14.9)$$

In cosine similarity,  $X_1$  and  $X_2$  are two  $n$ -dimensional vectors and it measures the angle between two vectors (thus called vector space model). Figure 14.2 shows the cosine similarity for different values of  $\theta$ .



**FIGURE 14.2** Cosine similarity of different values of  $\theta$ .

Consider movie ratings given by two customers on a 5-point scale for movies listed in Table 14.4.

**TABLE 14.4** Movie rating by two customers

Movie	Ghajini	Ra.One	Dilwale	3 Idiots	Fan	PK
Customer 1	5	1	1	5	1	5
Customer 2	1	5	5	1	5	1

Using Eq. (14.9) on the data shown in Table 14.4, we get

$$\text{Cosine similarity} = \frac{30}{8.832 \times 8.832} = 0.385$$

A lower cosine similarity indicates low similarity between two observations. Completely similar observations will have a cosine similarity value of 1.

#### 14.2.7 | Gower's Similarity Coefficient

The distance and similarity measures that we have discussed so far are valid either for quantitative data or qualitative data. Most data sets will have both qualitative and quantitative data. Gower's similarity coefficient (Gower, 1971) can be used when the data has both quantitative and qualitative data. Gower's coefficient between two  $n$ -dimensional observations  $i$  and  $j$  is given by

$$D_{ij} = \frac{\sum_{k=1}^n D_{ijk} W_{ijk}}{\sum_{k=1}^n W_{ijk}} \quad (14.10)$$

where  $D_{ijk}$  is the similarity between observations ( $i$  and  $j$ ) for  $k^{\text{th}}$  variable and  $W_{ijk}$  is a binary variable that captures whether the similarity calculation between observations is valid for  $k^{\text{th}}$  variable.  $W_{ijk} = 1$  if the similarity calculation is valid, that is, when both the values are known and 0 otherwise. That is,  $W_{ijk}$  takes value 1 only when comparisons are possible, otherwise zero (Gower, 1971).

The similarity  $D_{ijk}$  for ordinal and continuous variables is given by

$$D_{ijk} = 1 - \frac{|X_{ik} - X_{jk}|}{R_k} \quad (14.11)$$

Here  $X_{ik}$  is the value of observation  $i$  for variable  $k$  and  $X_{jk}$  is the value of observation  $j$  for variable  $k$ .  $R_k$  is the range of variable  $k$  in the data set. The value of  $D_{ijk}$  lies between 0 and 1.

For categorical variables

$$D_{ijk} = \begin{cases} 1 & \text{when } X_{ik} = X_{jk} \\ 0 & \text{otherwise} \end{cases} \quad (14.12)$$

$W_{ijk} = 1$  if  $X_{ik}$  and  $X_{jk}$  are known.

Gower (1971) proposed the following procedure for calculating  $D_{ijk}$  and  $W_{ijk}$  for binary variables.

1. If the variable takes value 1 for both observations then  $D_{ijk} = W_{ijk} = 1$ .
2. If either  $i^{\text{th}}$  or  $j^{\text{th}}$  observation takes value 1, then  $D_{ijk} = 0, W_{ijk} = 1$ .
3. When both  $i^{\text{th}}$  and  $j^{\text{th}}$  take values zero, then  $D_{ijk} = W_{ijk} = 0$ .

Table 14.5 shows 5 customers and their movie downloads from a portal. The data consists of genre of the movies, maximum rating given by the customer, and the marital status (code 1 implies married and 0 otherwise). For example, customer 1 downloaded 23 action, 5 romance, 15 comedy, and 0 Sci-fi movies and his maximum rating was 1.

**TABLE 14.5** Customer data on movie downloads

Customer	Number of Movies Downloaded Under Each Genre				Maximum Rating ( $k=5$ )	Marital Status ( $k=6$ )
	Action ( $k=1$ )	Romance ( $k=2$ )	Comedy ( $k=3$ )	Sci-fi ( $k=4$ )		
1	23	5	15	0	1	0
2	5	18	16	2	5	1
3	25	0	0	15	5	0
4	2	30	15	0	4	1
5	45	0	0	10	5	0

The Gower's similarity between customers 1 and 2 can be calculated as shown in Table 14.6.

**TABLE 14.6** Gower's distance calculations between customers 1 and 2

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	Sum
$D_{ijk}$	0.5814	0.5667	0.9375	0.8667	0.0000	0	2.952
$W_{ijk}$	1	1	1	1	1	1	6

The Gower's similarity coefficient between customers 1 and 2 is given by  $2.952/6 = 0.492$ .

### 14.3 | QUALITY AND OPTIMAL NUMBER OF CLUSTERS

One of the challenges in clustering is identification of the ideal number of clusters and the quality of clusters. According to Milligan (1996), the clustering techniques used should be able to identify the underlying clusters within the data. Milligan and Cooper (1985) analysed over 30 procedures for determining the optimal number of clusters and recommended the index proposed by Calinski and Harabasz (1974) which is given by

$$CH(k) = \frac{B(k)/k-1}{W(k)/(n-k)} \quad (14.13)$$

where  $CH(k)$  is the Calinski and Harabasz index with  $k$ -clusters ( $k > 1$ ),  $B(k)$  and  $W(k)$  are the between and within clusters sum of squared variations with  $k$  clusters. The optimal  $k$  is the one with maximum  $CH(k)$  value. Hartigan (1975) proposed the following statistic to decide the number of clusters:

$$H(k) = \left\{ \frac{W(k)}{W(k+1)} - 1 \right\} / (n - k - 1) \quad (14.14)$$

Hartigan suggested a cluster can be added as long as  $H(k)$  is more than 10 (Yan, 2005). Another popular measure used for deciding the number of clusters is the Silhouette statistic due to Kaufman and Rousseeuw (1990). Let  $a(i)$  be the average distance between an observation  $i$  and other points in the cluster to which observation  $i$  belongs. Let  $b(i)$  be the minimum average distance between observation  $i$  and observations in other clusters. Then the Silhouette statistic is defined by

$$S(i) = \left( \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \right) \quad (14.15)$$

A higher value of  $S(i)$  indicates better clustering.

## 14.4 | CLUSTERING ALGORITHMS

The distance measures discussed in Section 14.2 can be used to group the data into useful and differentiable groups. Clustering algorithms group data into finite number of mutually exclusive subsets. Assume that  $S$  is the set of all observations. Non-overlapping clustering algorithms attempt to create subsets ( $C_j$ ) of  $S$ , such that  $S = \bigcup_{j=1}^n C_j$ , and for any two subsets  $C_i$  and  $C_j$ ,  $C_i \cap C_j = \emptyset$  (null set) for  $i \neq j$ . There are many clustering algorithms that use different logic and distance measures. The objective of the clustering techniques is to create groups such that the variation within the group is minimized and variation between the groups is maximized. In this section, we will be discussing two most frequently used clustering methods, namely, K-means clustering and Hierarchical clustering. The following steps are followed in clustering algorithms:

1. Variable selection.
2. Deciding the distance/similarity measure for measuring distance/dissimilarity between the observations.
3. Deciding the number of clusters.
4. Validation of the clusters.

We will discuss these steps in detail next.

### 14.4.1 | Variable Selection

Ketchen and Shook (1996) suggest inductive, deductive, and cognitive approaches for variable selection. Inductive is basically an exploratory approach and starts with as many variables as possible. On the other hand, in deductive variable selection, suitability of the variable and theoretical basis influence selection of variables. Under cognitive variable selection, expert opinion plays a major role in variable selection (Ketchen and Shook, 1996).

### 14.4.2 | Deciding Distance/Similarity Measures

Choosing the right distance/similarity measure plays an important role in developing clusters. For example, Euclidean distance is valid only for variables under interval and ratio scales. However, for qualitative

variables, Euclidean distance is not valid. Similarity measures such as Jaccard coefficient should be used for binary variables. Cosine similarity can be used for both quantitative and qualitative variables. If the data has both qualitative and quantitative variables, then one may have to use measures such as Gower's distance.

#### 14.4.3 | Number of Clusters

Several approaches are available for deciding the number of clusters such as  $CH$  index [Eq. (14.13)], Hartigan statistic [Eq. (14.14)], Silhouette statistic [Eq. (14.15)], and elbow method in which the ideal number of clusters is given by the position of elbow in an  $L$ -shaped curve.

#### 14.4.4 | Cluster Validation

The clusters created should be validated for consistency using different algorithms to ensure that the clusters represent the structures that exist in the population. Halkidi *et al.* (2001) suggest the following measures to validate the clusters:

1. **Compactness:** Closeness of each member of a cluster which can be measured through variance.
2. **Separation:** Distance between different clusters.

### 14.5 | K-Means Clustering

K-means clustering is one of the frequently used clustering algorithms. It is a non-hierarchical clustering method in which the number of clusters ( $K$ ) is decided *a priori*. The observations in the sample are assigned to one of the clusters (say  $C_1, C_2, \dots, C_K$ ). The following steps are used in K-means clustering algorithm:

1. Choose  $K$  observations from the data that are likely to be in different clusters. There are many ways of choosing these initial  $K$  values; easiest approach is to chose observations that are farthest (in one of the parameters of the data).
2. The  $K$  observations chosen in step 1 are the centroids of those clusters.
3. For remaining observations, find the cluster closest to the centroid. Add the new observation (say observation  $j$ ) to the cluster with closest centroid. Adjust the centroid after adding a new observation to the cluster. The closest centroid is chosen based on an appropriate distance measure.
4. Repeat step 3 till all observations are assigned to a cluster.

Note that centroids keep moving when new observations are added; also observations may move to different clusters. An important aspect of K-means clustering is choosing the appropriate value of  $K$ . Initially the value of  $K$  is a guess; however, it can be decided based on several measures such as  $CH(K)$  index, Silhouette coefficient and elbow method.

K-means clustering is used to group 149 Bollywood movies (Data file: Bollywood Data Clustering.xls). The variables used along with descriptive statistics are given in Table 14.7.

**TABLE 14.7** Bollywood movie data for clustering

Variable	Minimum	Maximum	Mean	Standard Deviation
Box-office collection	0.01 (in crores)	735 (in crore)	55.67	94.49
Profit	-56.80	650.00	26.24	79.09
Earnings ratio (Ratio of box-office collection over budget)	0.01	9.17	1.77	1.84
Budget	1.8 (crores)	150 (crores)	29.43 (crores)	28.25 (crores)
YouTube views	4354	23171067	3337919.91	3504406.99
YouTube likes	1	101275	7877.54	12748.04
YouTube dislikes	1	11888	1207.82	1852.69

K-mean clustering output for  $K = 3$  using SPSS is shown in Tables 14.8–14.11.

**TABLE 14.8** Final cluster centres

	Cluster		
	1	2	3
Box_Office_Collection	306.10	72.89	32.42
Profit	215.801666	34.2598	10.9492
Earning_Ratio	3.40	2.08	1.53
Budget	90.3	38.6	21.5
Youtube_VIEWS	16399358	5506403	1542508
Youtube_Likes	52311	12857	2871
Youtube_Dislikes	7169	2068	448

Table 14.8 shows the mean values of variables in each cluster. Cluster centers in Table 14.8 help us to identify characteristics of various clusters. For example, the budget (and profit) of movies for cluster 1 is much higher than clusters 2 and 3. The number of YouTube likes is much higher for cluster 1 compared to clusters 2 and 3.

**TABLE 14.9** Distances between final cluster centres

Cluster	1	2	3
1		10893027.904	14856933.416
2	10893027.904		3963907.285
3	14856933.416	3963907.285	

Table 14.9 shows the Euclidean distance between centroids of the three clusters. Higher the distance, larger is the distance between the centroids. However, note that the scales of different variables are different and hence ideally the data should be normalized before performing cluster analysis. ANOVA for the variables used in clustering is shown in Table 14.10. Higher value of  $F$  indicates higher level of contribution of that variable in clustering.

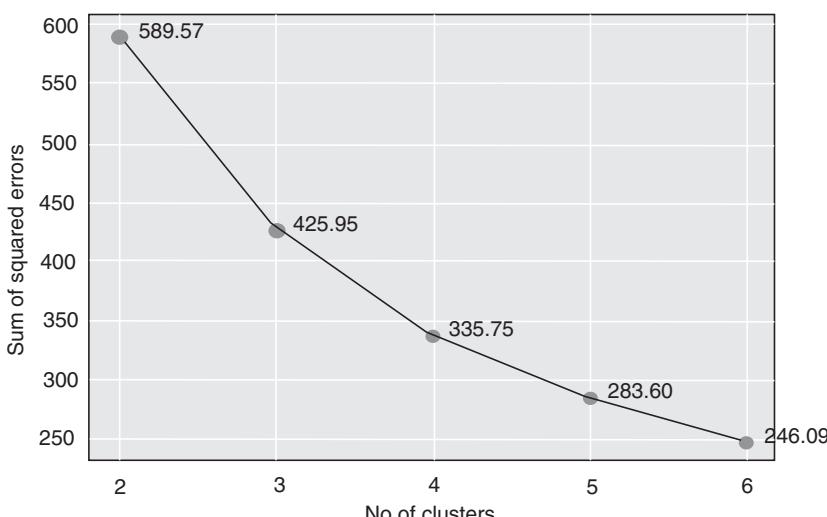
**TABLE 14.10** ANOVA for variables used in clustering<sup>a</sup>

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Box_Office_Collection	221299.22	2	6020.03	146	36.76	0.000
Profit	120704.64	2	4687.10	146	25.75	0.000
Earning_Ratio	12.98	2	3.24	146	4.00	0.020
Budget	16121.67	2	588.17	146	27.41	0.000
Youtube_VIEWS	775557272679141	2	1825027216893.38	146	424.95	0.000
Youtube_Likes	7709212606.51	2	59133256.50	146	130.37	0.000
Youtube_Dislikes	151589730.46	2	1402919.62	146	108.05	0.000

<sup>a</sup>The F-tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

The number of observations in three different clusters is shown in Table 14.11.

TABLE 14.11 Number of cases in each cluster		
Cluster	1	6
	2	45
	3	98
Valid	149	
Missing	0	

**FIGURE 14.3** Elbow curve for the Bollywood data.

The elbow curve for the clusters developed using the Bollywood data is shown in Figure 14.3 (obtained using R programming language). Based on the elbow curve we can conclude that the optimal number of clusters in this case is 3 (bend seems to appear when the number of clusters is 3).

## 14.6 | HIERARCHICAL CLUSTERING

Hierarchical clustering is a clustering algorithm which uses the following steps to develop clusters:

1. Start with each data point in a single cluster.
2. Find the data points with shortest distance (using an appropriate distance measure) and merge them to form a cluster.
3. Repeat step 2 until all data points are merged to form a single cluster.

The above procedure is called **agglomerative hierarchical cluster**. Agglomerative hierarchical clustering is explained by using the data in Table 14.12. There are 8 data points ( $D_1, D_2, \dots, D_8$ ) in Table 14.12 and distances between each of these observations are given.

**TABLE 14.12** Data points with distances

	$D_1$	$D_2$	$D_3$	$D_4$	$D_5$	$D_6$	$D_7$	$D_8$
$D_1$	0	0.45	0.36	0.71	1.00	0.27	0.38	0.21
$D_2$	0.45	0	0.19	0.36	0.54	0.30	0.91	0.76
$D_3$	0.36	0.19	0	0.87	0.54	0.72	0.28	0.64
$D_4$	0.71	0.36	0.87	0	0.34	0.51	0.43	0.72
$D_5$	1.00	0.54	0.54	0.34	0	0.65	0.57	0.41
$D_6$	0.27	0.30	0.72	0.51	0.65	0	0.33	0.68
$D_7$	0.38	0.91	0.28	0.43	0.57	0.33	0	0.44
$D_8$	0.21	0.76	0.64	0.72	0.41	0.68	0.44	0

In Table 14.12, the closest data points are  $D_2$  and  $D_3$ . So,  $D_2$  and  $D_3$  will be merged to form the first cluster, say  $C_1$ . Table 14.12 can be modified to denote this cluster as shown in Table 14.13. The distances between cluster  $C_1$  and other data points are calculated based on the maximum distance between the data points in the cluster and other data points which are not part of cluster  $C_1$ . For example, the distance between cluster  $C_1$  and data point  $D_1$  is 0.45 since the distance between  $D_1$  and  $D_2$  is 0.45 and the distance between  $D_1$  and  $D_3$  is 0.36. The process is repeated till all data points become part of one single cluster. The agglomerative hierarchical clustering can be represented using a tree-like structure called dendrogram (Figure 14.2).

**TABLE 14.13** Data points with distances after first cluster

	$D_1$	$C_1 = \{D_2, D_3\}$	$D_4$	$D_5$	$D_6$	$D_7$	$D_8$
$D_1$	0	0.45	0.71	1.00	0.27	0.38	0.21
$C_1 = \{D_2, D_3\}$	0.45	0	0.87	0.54	0.72	0.91	0.76
$D_4$	0.71	0.87	0	0.34	0.51	0.43	0.72
$D_5$	1.00	0.54	0.34	0	0.65	0.57	0.41
$D_6$	0.27	0.72	0.51	0.65	0	0.33	0.68
$D_7$	0.38	0.91	0.43	0.57	0.33	0	0.44
$D_8$	0.21	0.76	0.72	0.41	0.68	0.44	0

The next cluster will be  $D_1$  and  $D_8$ . Table 14.14 shows the table after cluster  $C_2 = \{D_1, D_8\}$ .

**TABLE 14.14** Data points with distances after second cluster

	$C_1 = \{D_2, D_3\}$	$D_4$	$D_5$	$D_6$	$D_7$	$C_2 = \{D_1, D_8\}$
$C_2 = \{D_1, D_8\}$	0.76	0.71	1.00	0.68	0.44	0
$C_1 = \{D_2, D_3\}$	0	0.87	0.54	0.72	0.91	0.76
$D_4$	0.87	0	0.34	0.51	0.43	0.72
$D_5$	0.54	0.34	0	0.65	0.57	1.00
$D_6$	0.72	0.51	0.65	0	0.33	0.68
$D_7$	0.91	0.43	0.57	0.33	0	0.44

The process is repeated till all observations become part of one cluster. For the Bollywood data, the SPSS output for hierarchical clustering is shown in Tables 14.15 and 14.16. For demonstration, only first 8 observations are used.

**TABLE 14.15** Squared Euclidean distance between cases

Case	Squared Euclidean Distance							
	1	2	3	4	5	6	7	8
1	0	2.51E + 13	5.04E + 13	6.73E + 13	7.41E + 13	7.43E + 13	8.05E + 13	8.3E + 13
2	2.51E + 13	0	4.37E + 12	1.02E + 13	1.29E + 13	1.3E + 13	1.57E + 13	1.68E + 13
3	5.04E + 13	4.37E + 12	0	1.22E + 12	2.27E + 12	2.32E + 12	3.52E + 12	4.05E + 12
4	6.73E + 13	1.02E + 13	1.22E + 12	0	1.61E + 11	1.73E + 11	5.92E + 11	8.2E + 11
5	7.41E + 13	1.29E + 13	2.27E + 12	1.61E + 11	0	2.53E + 08	1.36E + 11	2.55E + 11
6	7.43E + 13	1.3E + 13	2.32E + 12	1.73E + 11	2.53E + 08	0	1.25E + 11	2.4E + 11
7	8.05E + 13	1.57E + 13	3.52E + 12	5.92E + 11	1.36E + 11	1.25E + 11	0	1.87E + 10
8	8.3E + 13	1.68E + 13	4.05E + 12	8.2E + 11	2.55E + 11	2.4E + 11	1.87E + 10	0

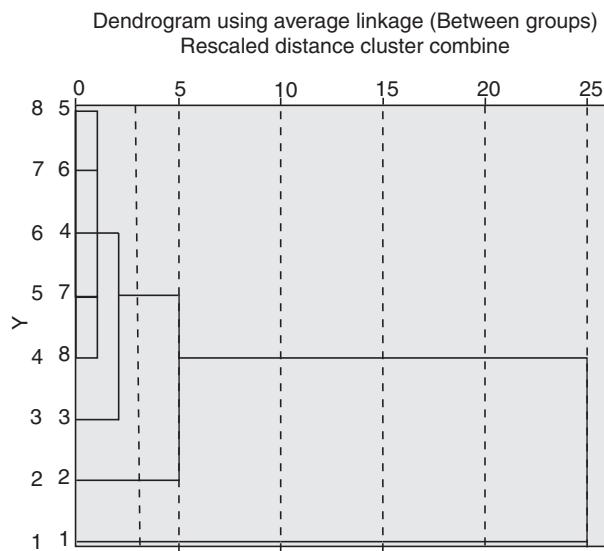
Table 14.15 provides squared Euclidean distance between the 8 cases considered for clustering. The schedule of the clustering is provided in Table 14.16. Initially all 8 observations are in individual clusters. In the first stage (first row of Table 14.16), cases 5 and 6 are merged to form a cluster since they have the minimum distance (Table 14.5) among all cases. Columns 5 and 6 report when the cluster has appeared

in the immediate past merging stage. For example, in stage 3, cases 4 and 5 are merged, but case 5 appears in stage 1 earlier. This is reflected in column 6. In stage 4, cases 4 and 7 are merged, and case 4 has already appeared in stage 3 (column 5) and case 7 has appeared in stage 2 (column 6). The last column indicates in which stage the cases merged in the current stage appear again. In Table 14.16, the coefficients denote the distance between the clusters.

**TABLE 14.16** Agglomeration schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	5	6	252516641.660	0	0	3
2	7	8	18725562050.840	0	0	4
3	4	5	166865087578.522	0	1	4
4	4	7	361200948423.055	3	2	5
5	3	4	2674635820607.503	0	4	6
6	2	3	12184893092907.690	0	5	7
7	1	2	64974081024938.125	0	6	0

The corresponding dendrogram is shown in Figure 14.4. Dendrogram is a pictorial representation of merging of various cases as the Euclidean distance is increased. The distance is rescaled to a scale between 0 and 25 in Figure 14.4. By drawing a vertical line at different values of re-scaled distance, one can identify the clusters. For example, assume that the vertical line is at a scaled distance of 20 (as shown in a dotted line in Figure 14.4). Then there will be 2 clusters  $C_1 = \{1\}$  and  $C_2 = \{2, 3, 4, 5, 6, 7, 8\}$ . If the vertical line is around the rescaled distance of 4, then the number of clusters will be 3 and they are  $C_1 = \{1\}$ ,  $C_2 = \{2\}$ ,  $C_3 = \{3, 4, 5, 6, 7, 8\}$ .



**FIGURE 14.4** Dendrogram for movie clustering.

There are large number of clustering algorithms reported in literature which differ mostly by distance and similarity measures used and logic used for deriving clusters. It is important that the

algorithm is able to identify the underlying structure of the clusters for any meaningful use. Once the clusters have been identified, the decision maker has to derive strategies for different clusters to maximize the value generated from each cluster.

## SUMMARY

1. Clustering is an unsupervised learning algorithms that divides the data set into mutually exclusive and exhaustive subsets (in non-overlapping clusters) that are homogeneous within the group and heterogeneous between the groups.
  2. Clustering is one of the frequently used techniques and practitioners first cluster the data and develop predictive models for each cluster for better management.
  3. Several distance measures such as Euclidian distance, Manhattan distance are used in clustering algorithms. Similarity, coefficients such as Jaccard coefficient and Gower's similarity are used depending on the data type.
  4. K-means clustering and Hierarchical clustering are two popular techniques used for clustering.
  5. One of the decisions to be taken during clustering is to decide on the number of cluster. Usually this is carried out using elbow curve. The cluster number at which the elbow (bend) occurs in the elbow curve is the optimal number of clusters.

## MULTIPLE CHOICE QUESTIONS

7. Which of the following is not used for calculating the optimal number of clusters?
  - (a) Elbow method
  - (b) Silhouette distance
  - (c) CH Index
  - (d) Jaccard coefficient
8. Distance between two clusters are measured between
  - (a) Centroids of two clusters
  - (b) Minimum distance between closest cases in the clusters
  - (c) Maximum distance between farthest cases in the clusters
  - (d) Average distance between all observations in each cluster

### EXERCISES

1. The movie ratings given by 4 customers ( $C_1$ ,  $C_2$ ,  $C_3$ , and  $C_4$ ) on five movies (A, B, C, D and E) are given in Table 14.17.

**TABLE 14.17** Movie ratings by customers

Movies → Customer ↓	A	B	C	D	E
$C_1$	4	1	3	2	4
$C_2$	3	2	4	4	2
$C_3$	3	3	4	4	4
$C_4$	3	3	2	3	3

Use cosine similarity to find among customers  $C_1$ ,  $C_2$ , and  $C_3$ , who is the closest to customer  $C_4$ .

2. An online store sells products under 8 categories labelled: A, B, ..., H. The past purchase details of 7 customers are given in Table 14.18.

**TABLE 14.18** Purchase history of products

Product → Customer ↓	A	B	C	D	E	F	G	H
$C_1$	1	0	0	1	1	0	1	1
$C_2$	1	0	1	1	1	1	0	0
$C_3$	0	1	1	0	0	0	1	1
$C_4$	1	0	0	1	1	0	0	0
$C_5$	1	1	1	0	0	0	0	1
$C_6$	0	0	1	1	0	0	1	0
$C_7$	1	1	0	0	0	1	1	1

where

$$a_{ij} = \begin{cases} 1, & \text{Customer } i \text{ purchased product } j \\ 0 & \text{otherwise} \end{cases}$$

Use Jaccard coefficient to find customer who is closest to customer  $C_1$ .

3. Customer feedbacks on 5 training programs (on a 5-point scale) by 6 customers are provided in Table 14.19.

**TABLE 14.19** Feedback on training programs

	$M_1$	$M_2$	$M_3$	$M_4$	$M_5$
$C_1$	2	4	2	4	3
$C_2$	4	3	2	4	5
$C_3$	1	2	3	2	4
$C_4$	4	4	2	4	3
$C_5$	2	1	2	2	3
$C_6$	2	1	1	4	4

- (a) Use cosine similarity to identify the customer who is closest to customer 1.  
 (b) Calculate correlation between different customers. Which customer has the highest correlation with customer 1?  
 (c) What is your conclusion based answers to questions (a) and (b)?
4. An online grocery store has captured amount spent per annum (in Indian rupees) by 20 customers on apparel and beauty and healthcare products. The data is shown in Table 14.20.

**TABLE 14.20** Amount spent by customers on apparel and beauty and healthcare products (in thousands of rupees)

Customer	Apparel	Beauty and Healthcare	Customer	Apparel	Beauty and Healthcare
1	21.1	0.7	11	5.2	16.2
2	15.23	5.5	12	14.2	2.9
3	5.22	18.6	13	4.4	19.4
4	31.1	1.8	14	4.25	15.5
5	6.12	21.5	15	22.3	0.9
6	14.5	8.2	16	7.9	18.8
7	8.5	16.2	17	13.4	4.2
8	26.5	2.2	18	30.6	1.9
9	4.34	17.7	19	14.4	6.28
10	13.75	7.3	20	6.25	9.98

- (a) Use K-means algorithm to find ideal number of clusters and cluster characteristics.  
 (b) Calculate the cluster centres of the clusters identified in (a).  
 (c) Calculate the distances between the clusters identified in (a).  
 (d) Use Hierarchical clustering to find the appropriate clusters for the data in Table 14.20.
5. The data set usedcars.xls has details of 1008 used cars along with the following variables: 1. Brand, 2. Car model, 3. Resale price, 4. Mileage, 5. Seat capacity, 6. Vehicle type, 7. Fuel type, 8. Transmission, 9. Parking sensor, 10. Airbag, 11. Cruise Control, 12. Keyless entry, 13. Alloy wheel, 14. ABS, 15. Climate control, 16. Rear AC vent and 17. Power Steering  
 (a) For the data set given, which distance measure is more appropriate?  
 (b) Use the distance measure identified in (a) and cluster the data. Identify the cluster characteristics.  
 (c) Use only the numerical variables (resale price, mileage, and capacity) in the data set and build clusters. Compare clusters developed in (b) and (c). Which clusters are better? Justify your answer.

## Markdown Optimization for an Indian Apparel Retailer<sup>1</sup>

The vision of WE SELL STYLE is to provide good quality merchandise at an affordable price to the Indian consumers.

— *Siddharth Sinha, Chief Executive Officer*

Siddharth Sinha, Chief Executive Officer of ‘WE SELL STYLE’<sup>2</sup>, started a discussion on pricing during end of the season sales with his team of managers. As it was one of the key challenges, he wanted the planning department to address it in the forthcoming season. He said:

While some apparel retailers position themselves as ‘good quality high price’ apparel retailers, we are providers of ‘good quality clothing at an affordable price’, hence our merchandise is very sharply priced, we need to be very astute in planning our markdowns during End of Season.

Markdown planning has been an important aspect of the apparel business. It is imperative to understand that the demand for fashion apparel is seasonal – affected by current fashion, variations in the seasons, festivals, and hence difficult to estimate. An apparel retailer could go off target – either by overestimating or underestimating the demand, with overestimating being prevalent. The ordering–manufacturing–stocking cycle is easily a 6-month cycle before the selling actually starts; with an expectation to improve sales year on year, the procurement team buys more, making an increase in the variety of colours and styles to offer more to the consumer. However, not all styles sell as expected, leaving higher than expected stocked inventory, which requires an impetus to sell. The impetus in the industry comes in the form of ‘end of season markdown’. Sinha summed up the problem as follows:

Our goal is to liquidate as much of our unsold inventory as possible, while keeping the markdowns at an optimal level. We don’t want to give too less, so as to be left with too much of an inventory, while at the same time being cautious of not offering too high to lose out on margins. At present, we are giving an overall markdown that is more than the industry standard for ‘everyday low price players’ like us.

WE SELL STYLE (WSS) sells apparel in the ‘Core’ and ‘Fashion’ categories. Core apparel is the vanilla, plain simple clothing, always in demand, while fashion apparel is high on design features such as prints, cuts, colours, and exhibit demand variability. The garments sold in core category are

<sup>1</sup> Deepak George, Karthik Kuram, Ramalakshmi Subramanian, Sumad Singh, and U Dinesh Kumar, Professor of DS & IS, prepared this case for class discussion. This case is not intended to serve as an endorsement, source of primary data, or to show effective or inefficient handling of decision or business processes.

Copyright © 2016 by the Indian Institute of Management Bangalore. No part of the publication may be reproduced or transmitted in any form or by any means – electronic, mechanical, photocopying, recording, or otherwise (including internet) – without the permission of Indian Institute of Management Bangalore. Reproduced with permission of IIM Bangalore.

<sup>2</sup> Name changed to maintain confidentiality.

## Continued...

on ‘planned markdown’ and are always sold in bundles – for example, buy 2 for INR 499 throughout the year (**Exhibit 1**).

The fashion garments come in a range of styles, often varying with seasons. Fashion garments experience high variability in sales, some styles and colours sell fast while others lag. At the end of a season, the left over fashion merchandise is put on markdown that often exceeds the planned markdown set for them. The focus of WSS is to reduce the unplanned markdowns to the extent possible; a process is in place to markdown styles by looking at their rate of sales and sell-through performance. However, the markdown given at WSS as a percent of sales is still above the industry standard.

## Apparel Retail Industry

### Market Outlook

A rising number of urban and small town consumers purchasing branded fashion led to the growth of organized apparel retail industry in India. Indian apparel industry has been showing a strong growth trend. According to a PricewaterhouseCoopers (PWC) 2015–16 Outlook for Retail and Consumer products sector, the apparel retail industry in India is forecast to have a value of \$15 billion in 2018, an increase of 91% since 2014, with the compound annual growth of the industry in 2014–18 predicted to be 17.5%<sup>3</sup> (**Exhibit 2**).

### Apparel Segments and Product Hierarchy

The apparel retail industry consists of the following segments – men’s wear, women’s wear, and kids’ wear. Apparel in each segment is further identified by a hierarchy of family, class, and brick (**Exhibits 3 and 4**). For example: Men’s wear is a segment; formal wear is a family under this segment; tops is a class of apparel; and shirts represent a brick. Other than the product hierarchy, apparels also have attributes. Attributes include brand, style, colour, and size. For example, Classic Polo is a brand; Polo Shirt with stripes is one of the styles; black is one of the colours; medium being one of the sizes.

Apparel stock keeping unit (SKU) is identified by brand, family, class, brick, style, colour, and size.

### Seasons

The apparel sale in India follows two seasons – Spring Summer and Festive Winter. The approximate periods of the seasons, interspersed with two end of season markdown periods are shown in **Table 14.21**.

### Apparel Lifecycle

Typical apparel lifecycle is shown in **Figure 14.5**. Four to five months before the launch of the season, an apparel retailer starts the lifecycle with the process of finalizing the styles to be launched in

<sup>3</sup> Source: 2015–16 outlook for retail and consumer products sector in Asia. [http://www.pwchk.com/webmedia/doc/635593364676310538\\_rc\\_outlook\\_201516.pdf](http://www.pwchk.com/webmedia/doc/635593364676310538_rc_outlook_201516.pdf)

## Case Study Continued...

the next season. Once the styles are decided, merchandising, the most critical activity starts. The retailer comes up with an assortment plan, that is, which styles should be introduced in which stores; the colour sets and size sets are also finalized. The assortment plan is used to create a buy plan, that is, what number to buy for each store considering the period of sales, store capacity, and expected rate of sales. The overall goal for a retailer operating in value merchandise is to sell more quantity and hence maximize profit in the season despite small margins. This requires an understanding of what merchandise is expected to sell in the period. The retailer must factor in the open inventory from the last season and then based on these data, determine the budget to be made available for procurement for the upcoming season, and hence determine the quantity to be procured.

After the planning process, the apparel is produced or ordered for production. Subsequently, floor plans and three-dimensional displays are charted out as a part of the visual merchandising process, to gear up for displaying the merchandise with launch of the season.

Retailer communicates the launch of the season with promotional and marketing activities. Once the selling/retailing starts, retailer monitors the sales vis-á-vis the planned forecast for every week of the season. If the sales underperform to the forecast, the retailer's planning and buying team decide to offer the merchandise at a price lower than the MRP. This is defined as markdown. Markdown done during the season is called in-season sale (referred to as ISS). Markdown offered at the end of a season is end-of-season sale (referred to as EOSS).

### About the Retailer

The WSS apparel retail chain<sup>4</sup> was set up in 2008, housing more than 100 brands. In 2015, they operated over 200 stores in all four regions of the country. They primarily focused on providing good quality fashion at a remarkably low price.

### Markdown

Decision on the percentage of markdown for EOSS is one of the most critical tasks for an apparel retailer. This activity starts months ahead of the EOSS. The product team and the planning team come up with an EOSS plan at the style level. In the decision process, procurement and planning team use their domain expertise and judge the performance of style using metrics such as rate of sales, full price sell-through, inventory left, and more. The key decision is to quantify the degree of non-performance of styles that did not sell as forecasted and by how much to markdown for the EOSS.

Often the procurement team does not favour heavy markdowns. Procurement teams are often criticized that they fall in love with what they buy, and continue believing that styles need to be on the shelves longer before their rate of sales picks up, and that they will make up to the

<sup>4</sup> This project used real client data for the analysis but owing to client confidentiality agreement, we have not mentioned actual client name.

**Continued...**

planned margins by the end of season. In contrast, the planners want to mark them down deep enough to recover costs and liquidate inventory, while settling for lower margins. So, this activity heavily depends on the teams involved, and consequently might not be the most optimal and objective. Another aspect that makes it sub-optimal is having same markdown plan across all the stores. These unplanned markdowns provide impetus to sales, but also result in depletion of margins (**Exhibit 5**).

### Scope of Changes

The management team started efforts to make improvements in having:

1. Markdowns based on groups of similar stores.
2. Markdown at brick level. In 2015, the EOSS planning was at the style level. With thousands of styles across different segments, the number of discount points became too high. Customer surveys showed that the high number of discount signage in the store led to clutter and confusion in the minds of the consumer. Management wanted to bring down the number by offering markdowns at a brick level.

### Markdown Optimization for Women's Ethnic Wear

Ramesh, the head of planning, had been trying to get over the 'one size fits all' approach of markdown planning for all stores. IT systems were capturing the sales data at 185 stores for all brands and bricks, along with promotions, and end of season markdown on apparel. He started looking at the data for the last festive winter EOSS period for women's ethnic wear. The data dictionary is provided in **Exhibit 6**.

Grouping stores that behave similarly during markdown period could enable the planning team to devise custom markdown plans for them. This idea of clustering was in contrast to the store categorization that was already being done using criteria such as sales per square feet, footfalls, sales conversion, catchment area, sell through, etc., to group stores for assortment planning, that is, which brand and styles should be stocked in what quantity and sold in which stores.

### Clustering

Sunil, VP of supply chain, who had heard about clustering to segment similar stores together said

I want to examine if the data we have on sales, markdown, cost of goods sold and store demographics can be used to cluster the stores, the immediate benefit could be for us in devising different promotion plans during EOSS for stores based on their cluster profiling.

**Continued...**

Sunil shared his idea with Ramesh.

I think we need a measure of how well the stores do in sales vis-à-vis the markdown given in EOSS period. Some stores may do exceedingly well in increasing sales when markdown is offered, while others may show no significant jump in sales, or possibly even a total slump in sales. For lack of a more comprehensive term, let us call it 'markdown sensitivity', where Markdown Sensitivity = Total Sales in EOSS (INR)/Total markdown given in EOSS (INR).

Ramesh was quick to formulate Analysis of Variance (ANOVA) to test the hypothesis that markdown sensitivity was not the same across stores during EOSS.

Confident with the results of hypothesis testing, Ramesh added:

We must also include a measure of profitability of store, normalized by the store area to have stores that are similar in profitability grouped closely and otherwise. Profit per Sq. Ft = Total Profit made during EOSS (in INR)/Total Area (in sq. feet).

Sunil before leaving, happy that he paved a way forward for clustering stores, commented

I would prefer including location as a measure in clustering also, to prefer closely located stores to be possibly clustered together.

### Demand Model Estimation

Sanjay, Head of Merchandizing and Shama, strategic advisor in the company were contemplating on the discussion from the morning in a separate meeting with their team. Shama opined.

Optimizing the markdown percentages requires determining the relationship between demand and markdown percentages. The relationship would mirror the typical demand – price curve.

Sanjay commented

The relationship does mirror the demand curve, we saw that when we ran a scatter plot between the two variables for women's ethnic wear category (**Exhibit 7**). However, if we try running a regression with units sold as dependent variable and percent discount given (markdown) during our last EOSS period, the fit statistics show a poor fit, and we know why that happened, don't we Ramya?

**Continued...**

Ramya, the Category Head, Women's Wear commented

Yes, we have a problem of specification bias. We are definitely missing out on other pertinent variables that explain the variation in demand.

We often see that there is a lag between the time we send out the marketing communication about EOSS and the surge in sales it causes, I can say that markdown we offered last week has a bearing on not only the sales of last week , but also on that of current week.

We see apparel demand also exhibits seasonality and trend, so we could model demand as a function of demand of previous weeks. Other factors like age of the merchandise, and the week being a markdown or non-markdown week need to be considered when we construct a mathematical model for demand estimation.

## Optimization

Ramya started off by getting the data in place to build a demand prediction model, while Shama's thoughts were trained on how this would fit into the objective of optimizing the markdown percentage during EOSS. In the EOSS season, a retailer wants to sell as much as possible while keeping as low markdowns as possible. The retailer came with a strategy of moving most of the apparel inventory to the other marts and stores under the same brand. With that, the pressure of keeping a low inventory at the end of season was somewhat relieved. The objective, therefore, was to earn as much revenue during the EOSS as possible.

The numbers from the last season's EOSS show that the retailer was able to dispose the merchandise that remained after EOSS in the partner stores at an approximate price of 40% of the original MRP.

Shama and Sunil in the meanwhile discussed the topic again over evening tea. Shama said

From what Ramya described, I see that the demand estimation model would definitely be a non-linear model. So, we are surely looking at a non-linear optimization model to solve the optimal discount percentages for the weeks of next EOSS.

Sunil responded

Indeed, but that should not be much of a challenge, we have non-linear solvers available aplenty to solve non-linear models. One needs to specify solving for a global optima and utilize 'multi-start' options available in the solver.

Continued...

### The Way Forward

Shama, Sunil, Sanjay, Ramya, and Ramesh came together and summed up the overall idea of the solution (**Exhibit 8**). The team agreed that the markdown optimization can be broken into three sub-problems:

1. Segment stores so that appropriate pricing strategy can be devised for each of the store clusters.
2. Develop a multivariate time series forecasting model to forecast demand.
3. Develop non-linear optimization model to identify optimal discounts that maximize revenue.



**EXHIBIT 1** 'Core' and 'Fashion' Merchandise. Core Merchandise – Plain men's shirts for everyday use, often sold in bundles of 2 or 3. Fashion Merchandise – salwars, tops, churidars, skirts, and denims with prints and cuts. The prints and cuts are inspired by what is expected to be liked in the season. Source: We sell style & Reliance Trends.

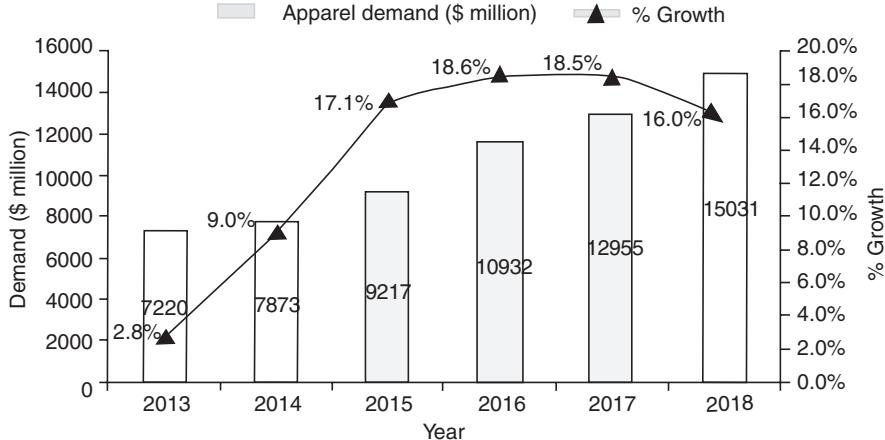
**EXHIBIT 2** Forecasted growth of Indian apparel retail business

Year	INR (billion)	% Growth
2013	916.7	15.9
2014	1058.7	15.5
2015	1229.3	16.1
2016	1352.7	10
2017	1496.6	10.6
2018	1695.5	13.3

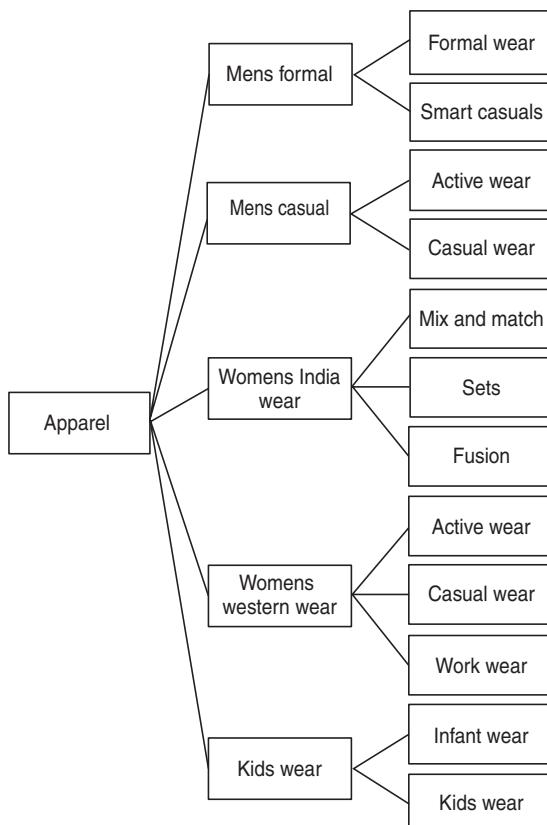
\*Source: Marketline industry profile of apparel retail in India, August 2014.

Case Study

Continued...



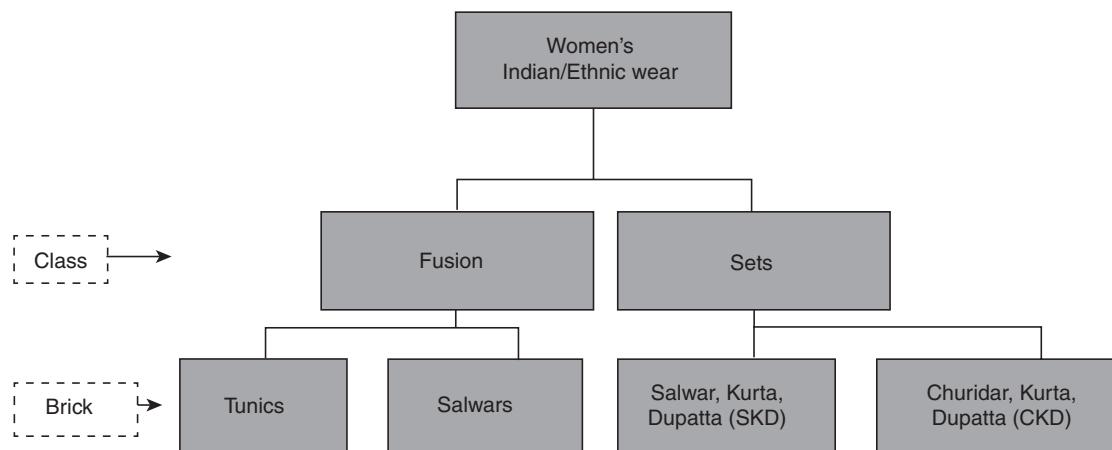
**EXHIBIT 2** \*Source: 2015–16 Outlook for retail and consumer products sector in Asia [http://www.pwchk.com/webmedia/doc/635593364676310538\\_rc\\_outlook\\_201516.pdf](http://www.pwchk.com/webmedia/doc/635593364676310538_rc_outlook_201516.pdf)



**EXHIBIT 3** Apparel segments and family. Source: Primary data collected from WE SELL STYLE.

Case Study

Continued...



**Exhibit 4** Example of apparel product hierarchy for women's ethnic wear<sup>5</sup> for brands – Brand-1 & Brand-2. Source: We sell style.

### Exhibit 5 An example of unplanned markdowns depleting margins while giving impetus to units sold

Order size of a style = 1000 units; MRP = INR 699;<sup>6</sup> average cost price = INR 400

Planned full price sell-through in the season = 75%, that is, 750 units planned to be sold at INR 699/unit

Merchandizing drew a plan for a planned sell-through of 25% at average discount of 30%, that is, 250 units planned to be sold at INR 489.3/unit

Target margin = INR 2,46,575

The plan of sale and margin realization, drawn at the beginning of selling period is as follows; last 4 weeks are EOSS weeks.

Week	Planned/Forecasted Sale (Units)	Selling Price	Margin Realization-Planned (INR)	Margin Realization-Planned (% age)	Cumulative Margin Realization-Planned (% age)
1	90	699	26910	11%	11%
2	120	699	35880	15%	25%
3	125	699	37375	15%	41%
4	135	699	40365	16%	57%

(Continued)

<sup>5</sup> Brand-1 and Brand-2 are two examples of brands sold by the retailer.

<sup>6</sup> 1 USD = INR 65.8 in 2016.

**Continued...**

Week	Planned/Forecasted Sale (Units)	Selling Price	Margin Realization-Planned (INR)	Margin Realization-Planned (% age)	Cumulative Margin Realization-Planned (% age)
5	100	699	29900	12%	69%
6	70	699	20930	8%	78%
7	60	699	17940	7%	85%
8	50	699	14950	6%	91%
9	70	489.3	6251	3%	93%
10	60	489.3	5358	2%	96%
11	60	489.3	5358	2%	98%
12	60	489.3	5358	2%	100%
Total	1000		2,46,575		

The actual sales turn out to be less than the planned/forecasted sale owing to seasonal variations, and hence margin realization is different from what is planned at the end of 8 weeks of regular selling. Using criteria such as rate of sale, inventory left, and full price sell-through, retailer decides to increase the markdown to 40% from the planned 30% for EOSS period. The actual sales and margin realization are shown in the following table.

Week	Actual Sale (Units)	Margin Realization-Actual (INR)	Margin Realization-Actual (% age)	Cumulative Margin Realization-Actual (% age)	% Discount Offered
1	40	11960	4.85%	4.85%	0%
2	50	14950	6.06%	10.91%	0%
3	60	17940	7.28%	18.19%	0%
4	80	23920	9.70%	27.89%	0%
5	60	17940	7.28%	35.17%	0%
6	60	17940	7.28%	42.44%	0%
7	50	14950	6.06%	48.50%	0%
8	40	11960	4.85%	53.35%	0%
9	90	1746	0.71%	54.06%	40%
10	80	1552	0.63%	54.69%	40%
11	70	1358	0.55%	55.24%	40%
12	60	1164	0.47%	55.72%	40%
Total	740	1,37,380			

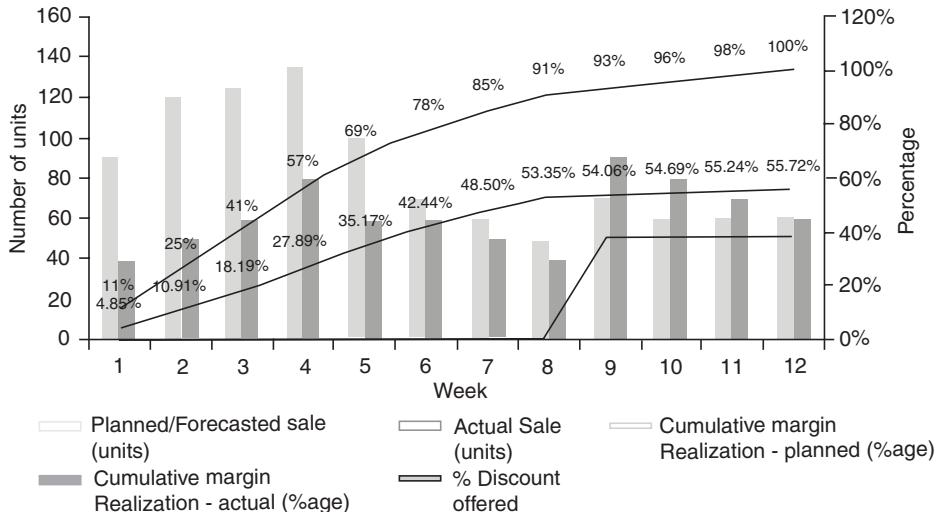
At the end of markdown period, the realized margin was 55.72% (INR 1,37,380) of the planned 100% (INR 2,46,575). It can be noted that:

- EOSS period experienced sales of 300 units, that is, 40% of sales (units) was achieved in last 4 weeks vs. 60% in first 8 weeks – see the actual sales increasing in the last 4 weeks in **Exhibit 5**.

## Case Study

**Continued...**

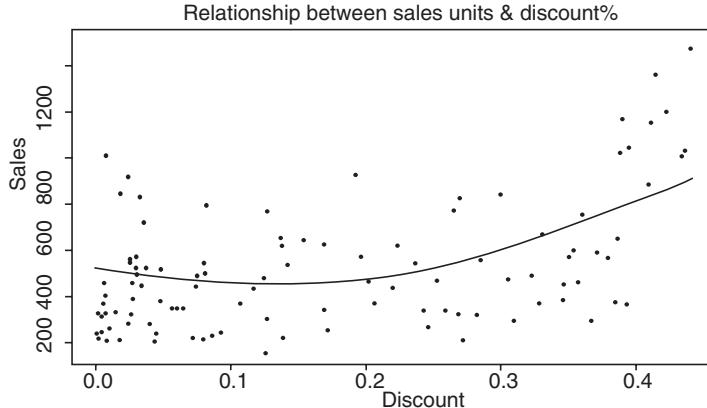
- The ending inventory as a result was at 260 units, which could have been far worse with 30% markdown.
2. Margin realized, however, rose only by 2.36% in EOSS period.



**EXHIBIT 6** Data dictionary of sales and markdown for women's ethnic wear across stores during last festive winter EOSS.

- (a) Store ID – Identifier for store
- (b) Store area – Area of store in square feet
- (c) Zone – East, west, north, and south zones based on location of store
- (d) Net sales in INR – Sales broken down by different brands under women's ethnic wear
- (e) Total discount in INR – Markdown broken down by different brands under women's ethnic wear
- f. Total cost of goods sold in INR – Cost of goods sold broken by different brands under women's ethnic wear

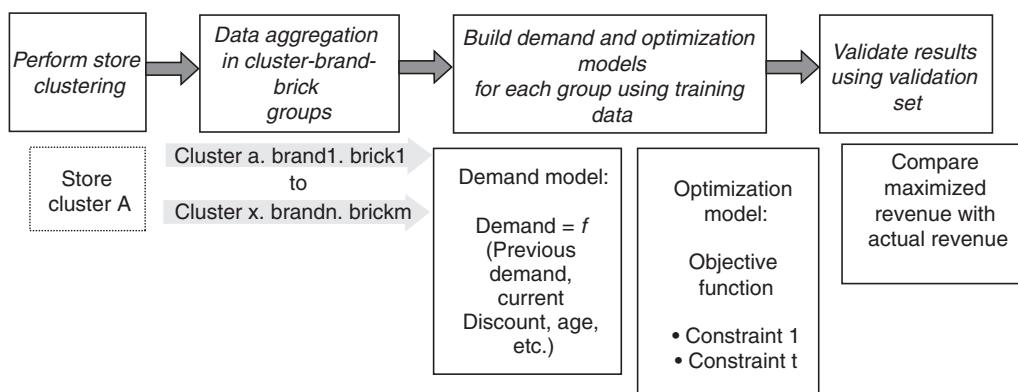
Source: WE SELL STYLE.



**EXHIBIT 7** Relationship between sales units and discount%. Source: Based on data from WE SELL STYLE.

Case Study

**Continued...**



**EXHIBIT 8** Solution approach. Source: WE SELL STYLE.

**TABLE 14.21** Apparel season in India

Season	Period	Weeks
Spring	Mid-January to Mid-April	16
Summer	Mid-April to June	10
Spring Summer EOSS	July to Mid-August	6
Festive	Mid-August to Mid-October	8
Winter	Mid-October to Mid-December	8
Festive Winter EOSS	Mid-December to Mid-January	4

Source: WE SELL STYLE.



**FIGURE 14.5** Apparel lifecycle. Source: WE SELL STYLE.

### CASE QUESTIONS

#### PART A – CLUSTERING

1. List and derive the metrics that can be used in ‘hierarchical clustering’ algorithms. (Note: Use the data in sheet “Clustering\_Raw\_data.xlsx”)
2. Do you find outliers in the derived data from question 1? If yes, how can the same be treated for use in cluster modeling?

**Continued...**

3. Develop a hierarchical clustering model with the modified data from question 2. How many clusters seem appropriate? Justify.
4. Develop a partition around medoids clustering model with the modified data from question 2. What are the advantages of using partitioning around medoids (PAM) over K-means? How do you decide on the appropriate number of clusters in this scenario?
5. Validate the goodness of resulting clusters from hierarchical and PAM models obtained in questions 3 and 4. Which is a better model as per validation measures?
6. Having selected the most appropriate clustering model from question 5, perform cluster profiling and list the unique characteristics of each cluster.

**PART B – TIME SERIES FORECASTING**

7. Conduct exploratory data analysis on ‘Cluster = 1, Brand = CRESCENT SET, Brick = CKD’ combination to identify the following relationships. Give a short description about the relationships observed.
  - (a) Relationship between Sales units (sales\_units) & Discount% (discount\_per)
  - (b) Relationship between Sales units and Net price (per\_unit\_netprice)
  - (c) Relationship between Sales units and Age (age)
- Notes:**
  - (1) Use data in the following csv files ‘1. CRESCENT SET.CKD.csv’ to answer this question.
  - (2) Variables names to be used are mentioned in brackets.
8. What is over-fitting and under-fitting in the context of regression models? What are the consequences of over-fitting?
9. Explain why we have to partition the time series data before building the forecasting model. Use data for ‘Cluster = 2, Brand = BLINK, Brick = HAREMS’ and partition this time series data as explained below.
  - (a) Consider all weeks until 51<sup>st</sup> week (including) of 2014 as training data.
  - (b) Consider weeks from 52<sup>nd</sup> week of 2014 to 3<sup>rd</sup> week of 2015 as test data.

**Notes:**

- (1) These 4 weeks are winter EOSS weeks.
- (2) User data in ‘2.BLINK.HAREMS.csv’ only for answering questions from now onwards.

10. Develop a time series forecast model using regression on the training data to forecast sales units for ‘Cluster = 2, Brand = BLINK, Brick = HAREMS’ combination using the below variables as predictors.
  - (a) Lag 1 (i.e. immediate previous week) of sales units
  - (b) Discount %
  - (c) Lag 1 (i.e. immediate previous week) of discount %
  - (d) Promotion week flag
  - (e) Age

Apply appropriate transformations and evaluate the model fit.
11. Perform checks to ensure that the model is valid and assumptions of regression are met. Conduct appropriate statistical test and back the findings by visual examination of relevant plots.
12. Based on the model result, explain the following:
  - (a) How is this forecasting model able to account for trend and seasonality?
  - (b) What is price elasticity and determine the price elasticity value (a proxy representing price elasticity is enough) from the model output?

**Continued...**

- (c) How do you interpret the coefficient of promotion week flag variable?
13. How do you check if the forecasting model is able to explain most of the important features of the time series? Explain white noise in the context of time series.
14. Using the forecast model built, generate sales units forecast for test period (52<sup>nd</sup> week of 2014 to 3<sup>rd</sup> week of 2015).  
 (a) Assess the forecast model accuracy on the test time period which is not used for modeling by calculating MAPE for the test period(d)

**Note:**

In 'Forecast\_test\_week\_predictor\_input.xlsx' sheet 'Input data for forecasting' contains the data required for forecasting sales and sheet 'Actual Sales' contains the actual sales

## PART C – OPTIMIZATION

15. Formulate an optimization model and solve it to determine the optimal discount % to be given for 'Cluster = 2, Brand = BLINK, Brick = HAREMS' combination for each of the 4 weeks of EOSS.  
 (a) Objective function: Maximize the total revenue generated during EOSS.  
 (i) Total revenue is defined as the revenue generated during 4 weeks of EOSS and revenue from the left over inventory after EOSS. Assume the residual inventory left over even after EOSS is liquidated by giving a flat discount of 60%.
- (b) Constraints  
 (i) Sales units in a week should be less than equal to starting inventory of the same week.  
 (ii) Relationship between demand (sales unit) for a week and predictors should follow the relationship identified in the forecasting regression model.  
 (iii) Discount for a week should be greater than equal to previous week's discount and discount should vary between 10% and 60% only.  
 (iv) Sales and inventory values for a week should be  $\geq 0$ .  
 (v) Inventory should be updated on basis of the sales for a week.
- (c) Inputs required  
 (i) Inventory at the beginning of EOSS = 2,476  
 (ii) Age of the brick at the beginning of EOSS = 96 weeks  
 (iii) Previous week discount = 57.9%  
 (iv) Previous week sale = 48  
 (v) MRP of the brick = INR 606
16. What are the weekly forecasted sales units if the optimal discounts identified are implemented for the EOSS?
17. We know that actual revenue realized by the retailer for 'Cluster = 2, Brand = BLINK, Brick = HAREMS' combination during the 4 weeks of EOSS is INR<sup>7</sup> 41,320. Then, what is the incremental lift in revenue the retailer would have achieved in these 4 weeks if he/she implemented our analytics solution instead?

<sup>7</sup> 1 USD = INR 65.8 in May 2016.

**REFERENCES**

1. Calinski T and Harabasz (1974), "A Dentrie Method for Cluster Analysis", *Communications in Statistics*, **3**, 1–27.
2. Gower J C (1971), "A General Coefficient of Similarity and Some of Its Properties", *Biometrics*, **27**(4), 857–871.
3. Halkidi M, Batistakis Y, and Vazirgiannis M (2001) "On Clustering Validation Techniques", *Journal of Intelligent Information Systems*, **17**(2/3), 107–145.
4. Hartigan J (1974), "Clustering Algorithms", Wiley, 1975.
5. Kaufman L and Rousseeuw P J (1990), "Finding Groups in Data: An Introduction to Cluster Analysis", Wiley Interscience, New York.
6. Ketchen D J and Shook C L (1996), "The Application of Clustering Analysis in Strategic Management Research: An Analysis and Critique", *Strategic Management Journal*, **17**(6), 441–458.
7. Milligan G W and Cooper M C (1985), "An Examination of Procedures for Determining the Number of Clusters in a Data Set, *Pschometrika*, **50**, 159–179.
8. Milligan G W (1996), "Clustering Validation: Results and Implications for Applied Analyses", *Clustering and Classification* (Eds. Arabie P., Hubert L J and Soete G De), World Scientific, Singapore.
9. Real R and Vargas J M (1996), "The Probabilistic Basis for Jaccard's Index of Similarity", *Systemic Biology*, **45**(3), 380–385.
10. Yan M (2005), "Methods of Determining the Number of Clusters in a Data Set and a New Clustering Criterion", *Ph.D. Thesis*, Virginia Polytechnique Institute and State University.



# 15

# Prescriptive Analytics

“Wherever you see a successful business, someone once made a courageous decision.”

—Peter Drucker

## LEARNING OBJECTIVES

- LO 15-1** Learn how prescriptive analytics techniques are used by organizations.
- LO 15-2** Understand complexities associated with prescriptive analytics problems and non-availability of efficient algorithms for many prescriptive analytics problems.
- LO 15-3** Learn various operations research techniques such as linear programming, integer programming, and goal programming and formulation of a problem as linear/integer/goal programming problem.
- LO 15-4** Learn the concept of shadow price and reduced cost in linear programming and how it can be used for decision making.
- LO 15-5** Understand integer programming problems and multi-objective problems.

## PREScriptive ANALYTICS

IMPORTANT

*Prescriptive analytics provides the optimal solution (or the best action) to a problem. Traditionally, Operations Research (OR) techniques are used for finding the optimal solution to a problem. Many machine learning algorithms use optimization techniques such as gradient descent while solving a problem.*

### 15.1 | INTRODUCTION TO PREScriptive ANALYTICS

Prescriptive analytics is the frontier of business analytics which is used for identifying the best action (or optimal solution) for a problem. Operations Research (OR) techniques are frequently used for finding optimal solution to a problem. Descriptive analytics analyses the past data to understand trends present in the data (answers to the question: What happened?). Predictive analytics techniques will predict what will happen to key performance indicators in the future? Prescriptive analytics assists the decision

maker to identify the best action (optimal solution), given the problem context. That is, prescriptive analytics, as the name suggests, prescribes the best solution or decision/action for the problem. Note that decisions or actions can be derived based on descriptive and predictive analytics as well. For example, using predictive analytics, retailer such as Amazon and Flipkart can predict what a customer is likely to buy in the future and design product recommendations. The difference between decisions arrived using descriptive/predictive analytics and prescriptive analytics is that prescriptive analytics algorithm tries to arrive at the best decision (optimal solution) based on an objective function (sometimes more than one objective functions) and a list of constraints.

Operations research techniques such as linear programming, integer programming, goal programming, non-linear programming, and meta-heuristics are used for prescribing optimal solution to a problem. A few big data problems originated from optimization problems. For example, travelling salesman problem (TSP) is one of the most difficult problem which is encountered by organizations such as online retailers, logistic service providers, and even electronic parts manufacturers. In a TSP, a salesman visits  $n$  cities from a starting point exactly once and returns to the starting point; the objective of the TSP is to minimize the total distance (or time taken) to complete the tour. There are many optimization problems which are classified as NP-hard problems for which there exists no efficient algorithm and exhaustive search is impossible (such as TSP, bin-packing problem, facility location problem, etc.). NP (Non-deterministic Polynomial time) problems are a class of problems for which the optimal solution can be found by non-deterministic Turing machine in polynomial time. For example, a travelling salesman problem (TSP) with 20 cities has a solution space of 20 factorials. If a computer can evaluate one million routes per second, it will take approximately 77,147 years to exhaustively search all 20 factorial routes. In absence of any efficient algorithms it is difficult to solve such problems.

Traditionally, big data problems are identified through volume, velocity, variety, and veracity of the data. However, there are many problems such as TSP for which there exists no efficient algorithms, making it one of the difficult problems to solve. At the same time many organizations deal with TSP frequently.

Prescriptive analytics models attempt to solve complex optimization problems of modern era. Modern electronic retailers (E-tailers) deliver millions of orders daily which means solving large number of travelling salesman problems. E-tailers have service-level agreement (SLA), that is, whenever a customer places order for a product on their website a delivery date is provided to the customer; not meeting this delivery date will result in customer dissatisfaction. For example, Ocado the online super market receives over 250,000 orders in a week (Armstrong, 2016). Indian online grocery store bigbasket.com delivers close to 300,000 orders every week (Abraham, 2016) and delivery promise within a time window is made to the customers. Each of these customer order may contain several items which are stored at different locations in Ocado's/bigbasket's warehouse. Collecting items in a specific customer order from different locations of the warehouse is a TSP. That is, Ocado and bigbasket have to solve several thousand TSP on a weekly basis which can be achieved only through machine learning algorithms. Most machine learning algorithms solve an optimization problem internally, thus prescriptive analytics has become an important part of analytics based solutions.

## 15.2 | LINEAR PROGRAMMING

Linear programming is one of the important techniques in operations research and prescriptive analytics. Linear programming is used when the objective function and the constraints of the problem can be

expressed as linear equation of decision variables. The use of linear programming dates back to the World War II during which manpower and logistics related problems were encountered by the US military and attempts to solve these problems were carried using linear programming techniques (Dantzig, 1963). Immediately after the World War II, several commercial applications of linear programming were identified which triggered further development of the field and solution approaches. The simplex algorithm proposed by Dantzig in 1947 (Dantzig, 1963) is an efficient algorithm which is one of the popular algorithms used for solving large-scale linear programming problems. Simplex means an  $n$ -dimensional polytope. In linear programming, the feasible region, if exists, will be a polytope. Large number of problems today can be either formulated as linear programming or integer programming problem that have significant influence on the profitability of organizations. Modern-day problems can have several million (if not billion) decision variables and are solved using sophisticated software tools such as IBM CPLEX and FICO Xpress. Few complex prescriptive analytics problems are listed below:

1. Travelling salesman problem which occurs in industries such as e-commerce (delivery of goods to customers), electronics parts manufacturing (movement of robot during manufacturing), and logistics service providers. Even honey bees encounter TSP (Morell, 2012); each bee (travelling salesman) has to collect honey from several flowers (cities).
2. Airlines encounter several prescriptive analytics problems such as tail assignment and maintenance optimization which for large airlines can result in several million decision variables. Each aircraft is uniquely identified through the number in its tail (and thus tail assignment problem). Tail assignment problem involves assigning an aircraft to a flight subject to several constraints (such as connectivity, minimum ground time, runway restrictions, etc.). The number of decision variables can run into several millions even for a mid-size airline with 100+ aircrafts. Aircrafts go through several maintenance checks such as weekly check, A check, B check, C check, and D check. These checks have to be conducted at specific intervals to maintain the airworthiness of the aircraft. Scheduling maintenance of aircraft for various maintenance checks in an optimal manner is a complex optimization problem leading to millions of complex constraints (Barnhart *et al.*, 2013).
3. Water distribution system in every city requires optimal design of water pipe network which is a complex prescriptive analytics problem.
4. Finding optimal location for utilities such as water tanks, fire stations, mobile towers, police stations are complex decision problems which are solved using prescriptive analytics techniques (Dicken, 1977).
5. Retail stores have to solve problems such as assortment planning (range of products that the retail store would like to sell) and shelf space allocation (allocation of shelf space across categories of products sold by the retailer). The number of individual stock keeping units (SKUs) sold by a brick-and-mortar retail store can be several thousands. Allocating space for individual SKUs will be complex and requires use of both predictive and prescriptive analytics.
6. Markdown optimization is a problem that is encountered by many retailers selling short-shelf-life products such as apparel and other fashion goods (bags, footwear, etc.). Retailers have to optimize the markdown (price reduction) at the end of the season to maximize the profit. This involves sales forecasting under different discounts such as 10%, 20%, etc. and the retailer has

to decide the optimal markdown which requires integration of both predictive and prescriptive analytics.

7. Scheduling such as nurse scheduling and airline crew can be very complex and requires prescriptive analytics to solve it. Hospital with 250 beds may have more than 1500 nurses of varying skills and usually work in 3 shifts. The hospital has to schedule nurses during each shift satisfying several constraint to optimize an objective function.

The aforementioned problems are few examples of thousands of business problems that are solved using prescriptive analytics techniques. A selected list of prescriptive analytics problems are discussed below:

1. **Product Mix Problem:** Frequently encountered problem across many industries in which the decision maker has to decide the number of different products to be produced using common resources. For example, companies such as Kellogg produce several products (variants of breakfast cereal) using common resources. Given the market demand, they would like to optimize the individual products to be manufactured during a given period under several resource constraints.
2. **Blending Problem:** Blending refers to mixing various ingredients to optimize yield of various end products and objective (such as profit). For example, petroleum refineries use various types of crude oil mixes (crude basket) to build various petroleum products (such as petrol, diesel, naphtha, etc). Refineries would like to optimize yield under constraints such as availability of different types of crude oil.
3. **Cutting Stock Problem:** Many industries (for example, paper, steel, glass, wood, etc.) have to minimize the waste generated due to cutting the original products into end products. For example, paper manufactures have to cut the original deckle size to end products such as A4, A3 size, etc. This process of cutting the original size to end products will result in cutting loss and the objective is to minimize the loss. The problem is an NP-hard problem and can get complex depending on the number of end products generated from the original product.
4. **Transportation Problem:** In a transportation problem the objective is to minimize the cost of transportation of goods from multiple origins (production centers) to several destinations (consumption centers).
5. **Assignment Problem:** The objective of the assignment problem is to assign task among agents (a task is assigned to one agent) that minimize the total assignment cost.
6. **Location Problem:** It involves optimal location of facilities to optimize objectives such as median (total distance) or minimize the maximum distance (distance between various residential areas and fire stations).
7. **Set Covering Problem:** Set covering problems are set of problems where the objective is to identify a subset from a set of elements such that the subset will cover the entire problem under given conditions. For example, assume that a city is divided into 10 regions, that is,  $S = \{1, 2, \dots, 10\}$  and the objective is to minimize the number fire stations to be located among these locations such that any region can be reached within 10 minutes. Set covering is again a NP-hard problem.

Note that the problems described above are not necessarily pure linear programming; few of them are modelled as integer programming problems.

### 15.3 | LINEAR PROGRAMMING (LP) MODEL BUILDING

First stage in LP is formulating the problem as LP problem. The following steps are used in formulating a problem as linear programming problem (LPP):

- Identification of decision variables:** Given a problem, we have to first identify the decisions to be taken by the decision maker. The decisions to be taken are expressed through decision variables.
- Identify the objective function:** The primary goal of the decision maker is expressed through the objective function which is a linear function of decision variables. The goal is either to minimize or maximize the objective function value.
- Identify constraints:** Constraints are restrictions such as availability of resources that a linear programming problem should satisfy.
- Identify implicit constraints:** Implicit constraints are conditions that the model has to satisfy, for example, the number of products to be produced cannot take negative values, and thus this variable can take only non-negative values. Also all variables need to be non-negative in simplex algorithm.
- Solve the problem:** Once the objective function and constraints are identified, the problems can be solved using algorithms such as simplex algorithm and interior point algorithm.
- Perform sensitivity analysis:** The values of objective function coefficients and resource availability may change due to several factors such as market conditions. It is important to understand the impact of the changes in objective function coefficient and resource availability on the optimal solution; this is achieved through sensitivity analysis.

#### EXAMPLE 15.1

#### Product Mix

Appukuttan Menon is the co-founder and CEO of Appukuttan Halva (AH) with headquarters in Kuttanad, Kerala. AH manufactures two types of halva: (a) Death by Halva (DH) and (b) Travancore Halva (TH). The main ingredients of halva are: (a) corn flour, (b) sugar, (c) fruit and nut, and (d) ghee. The quantity of each ingredient required for the two types of halva for every one kilogram is given in Table 15.1.

TABLE 15.1 Ingredients required for 1 kg of halva

Halva Type	Ingredients (in grams) Required for 1 kg of Halva			
	Corn Flour	Sugar	Fruit and Nut	Ghee
Death by Halva (DH)	500	750	150	200
Travancore Halva (TH)	500	625	100	300

The profit from DH and TH per kilogram are INR 45 and 50, respectively. The maximum daily demand for DH and TH are 50 kg and 20 kg, respectively. Appukuttan Menon is a big fan of the Japanese lean management concept and used JIT procurement. All the ingredients necessary for the daily production are delivered on the day of production at 6.00 am and AH maintained no safety stock. The DH and TH are delivered to the customers (local retail stores in Kuttanad) from 12.00 Noon onwards every day. The suppliers of the ingredients are located in Coimbatore, which is about 300 km from Kuttanad.

Due to some supply chain disturbance, suppliers of AH have informed Appukuttan Menon that they will be unable to supply the raw material on 25<sup>th</sup> January 2017; however, the supply will be restored from 26<sup>th</sup> January onwards. To manage the supply of halva on 25<sup>th</sup> January 2017, Appukuttan Menon decided to procure the ingredients locally. His procurement manager George Varghese informed him that since AH uses specific brands of the ingredients, the availability of raw material is limited in the local market and is shown in Table 15.2.

**TABLE 15.2** Availability (in grams) of ingredients in local market

Corn Flour	Sugar	Fruit and Nut	Ghee
20,000	42,000	10,400	9,600

Use linear programming to find the optimal product mix for 25<sup>th</sup> January for AH that will maximize the profit for AH? Assume that there will be no change in the profit of DH and AH due to procurement of ingredients from the local market.

### Solution:

---

**STEP 1** *Identification of the decision variables*

---

In this case, AH has to decide the quantity (in kilograms) of DH and TH to be produced. Let

$X_1$  = Quantity (in kilogram) of DH to be produced

$X_2$  = Quantity (in kilogram) of TH to be produced

---

**STEP 2** *Identification of the objective function*

---

The objective is to maximize the profit. The profit on DH per kg is 45 and the profit on TH per kg is 50. The objective function is

$$\text{Maximize } 45X_1 + 50X_2$$


---

**STEP 3** *Identify the constraints*

---

In this example, the constraints are availability of various ingredients.

**Constraint for corn flour:** 20,000 grams of corn flour is available. Each kg of AH requires 500 grams of corn flour and each kg of TH 500 grams of corn flour. Thus, the corresponding constraint is

$$500 X_1 + 500 X_2 \leq 20,000$$

**Constraint for sugar:** 42,000 kg of sugar is available. Each kg of DH requires 750 grams of sugar and each kg of TH 625 grams of sugar. Thus the corresponding constraint is

$$750 X_1 + 625 X_2 \leq 42,000$$

**Constraint for fruit and nut:** 10,400 kg of fruit and nut is available. Each kg of DH requires 150 grams of fruit and nut and each kg of TH requires 100 grams of fruit and nut. Thus the corresponding constraint is

$$150X_1 + 100X_2 \leq 10,400$$

**Constraint for ghee:** 9,600 kg of ghee is available. Each kg of DH requires 200 grams of ghee and each kg of TH requires 300 grams of ghee. Thus the corresponding constraint is

$$200X_1 + 300X_2 \leq 9,600$$

**Maximum demand constraint:** The maximum daily demand for DH and TH are 50 kg and 20 kg, respectively, which can be written as  $X_1 \leq 50$  and  $X_2 \leq 20$ .

#### STEP 4 Identify implicit constraints

In this case the implicit constraints are the quantity of DH and TH which cannot be negative. Thus the values  $X_1$  and  $X_2$  are non-negative. That is,  $X_1 \geq 0$  and  $X_2 \geq 0$ .

The complete LP formulation is

$$\text{Maximize } 45X_1 + 50X_2$$

subject to constraints

$$500X_1 + 500X_2 \leq 20,000$$

$$750X_1 + 625X_2 \leq 42,000$$

$$150X_1 + 100X_2 \leq 10,400$$

$$200X_1 + 300X_2 \leq 9,600$$

$$X_1 \leq 50$$

$$X_2 \leq 20$$

$$X_1 \geq 0 \text{ and } X_2 \geq 0$$

**Solution to LP Problem:** A linear programming model has the following properties:

1. The feasible region (set of solutions that satisfy all constraints) is a convex set. Convex set is a set such that for any two points  $(X_1, Y_1)$  and  $(X_2, Y_2)$  of the set,

a straight line joining these two points is also a part of the set. In case of linear programming with many variables, the feasible region is a **convex polytope**.

2. The optimal solution to a linear programming problem (LPP) is an extreme point. Extreme points are the vertices of a convex polytope and are also known as corner points.
3. Each LP problem has finite number of corner points.
4. To find the optimal solution, we have to search the corner points of the feasible region.

Simplex method proposed by Dantzig (1963) starts with a corner point and moves along the edge of the convex polytope to identify the next corner point which optimizes (maximize or minimize) the objective function value. During each iteration the algorithm moves from one corner point to another corner point in the direction of optimal solution and stops once an optimal solution is found. LP problem can be solved using the Solver Add-In in Microsoft Excel®. The Solver output for Example 15.1 is shown in Table 15.3.

**TABLE 15.3** Optimal solution to Example 15.1 using Solver

$X_1$	$X_2$	Objective Function Value
24	16	1880
<b>Constraints</b>		
LHS		RHS
20000	20000	Corn flour constraint
28000	42000	Sugar constraint
5200	10400	Fruit and nut constraint
9600	9600	Ghee constraint
24	50	Maximum demand for DH
16	20	Maximum demand for TH

The optimal solution (optimal quantity of DH and TH to be produced) to the problem is  $X_1 = 24$  and  $X_2 = 16$ . The maximum profit is 1880.

## 15.4 | LINEAR PROGRAMMING PROBLEM (LPP) TERMINOLOGIES

The following terminologies are used in LPP.

1. **Feasible region:** Feasible region is the set of solutions to the problem that satisfies all the constraints. The feasible region may be **bounded** or **unbounded**.
2. **Binding constraints:** Binding constraints are constraints for which the left-hand side (LHS) and right-hand side (RHS) of the constraint at the optimal solution are equal. For example, in

Table 15.3, the LHS and RHS are same for corn flour and ghee constraints, thus these are two binding constraints. An important property of binding constraint is that changing the RHS value (within a range) of the binding constraint will change the values of optimal solution.

3. **Non-binding constraint:** Non-binding constraints are constraints for which the values of LHS and RHS of the constraint are not same in the optimal solution. In Example 15.1, for the constraints sugar and fruit and nut, the values of LHS and RHS are not the same (Table 15.3). For a non-binding constraint, changing the value of the RHS (within a range) will not change the optimal solution.
4. **Slack variable:** In a less than or equal to constraint (in the current example, all constraints are less than or equal to constraints), slack variable is the unused resource. In general, slack variable is a variable which when added to LHS of a less than or equal to constraint will make it to an equality constraint. In Table 15.3, adding 5200 to LHS of the fruit and nut constraint will make it an equality constraint. Thus, the value of slack variable for the fruit and nut constraint is 5200.
5. **Surplus variable:** In a greater than or equal to constraint, surplus variable is the excess resource used (note that resource is a generic term that we are using, we have to interpret the excess based on the context). In general, surplus variable is a variable which when subtracted from LHS of a greater than or equal to constraint will make constraint into an equality constraint.
6. **Basic variable:** A decision variable with non-zero value in the optimal solution is called a basic variable. Set of basic variables is called the basis (term borrowed from matrix algebra).
7. **Non-basic variable:** A decision variable which has a value of zero in the optimal solution is called a non-basic variable.
8. **Shadow price:** Shadow price is the marginal value of a resource, that is, it is the value by which the objective function value changes for unit change in the RHS of a constraint.
9. **Reduced cost:** In the optimal solution to an LPP, it is possible that a decision variable is a non-basic variable (that is, it takes value 0). Reduced cost is the value by which the objective function value will deteriorate when a non-basic variable is forced to become a basic variable (that is, when the value of the non-basic variable is changed from 0 to 1).

## 15.5 | ASSUMPTIONS OF LINEAR PROGRAMMING

LPP models are built under the following assumptions:

1. **Additivity:** The value of the objective function is assumed to be sum of contribution of each decision variable value. That is, we assume that the contribution of each decision variable is independent. This assumption may not be satisfied in many cases. For example, assume that a marketing team is trying to find an optimal advertisement mix that will have maximum reach at minimum cost. The various channels for advertisement are: television, newspapers, radio, social media, etc. However, all these channels are not independent since there will be an overlap between channels. That is, if 5 million customers can be reached using television and 3 customers using radio, we cannot conclude that if we use both channels then 8 million customers can be reached, since there will be common customers between television and radio.

2. **Proportionality:** In LPP we assume that a change in a variable will result in proportionate change in the objective function value. This may not be true in all cases. For example a manufacturer may provide discount based on the quantity of purchase (quantity discounts) and thus the contribution may follow a step function.
3. **Divisibility:** An important assumption in LPP is that a decision variable can be broken into non-integer values. However, this assumption is relaxed in the case of integer programming problems.
4. **Certainty:** We assume that all the model parameters (objective function coefficient values and RHS of constraints) are known with certainty.

## 15.6 | SENSITIVITY ANALYSIS IN LPP

An important step in LPP is sensitivity analysis. Although we assume that the parameter values are known with certainty, it is likely to change due to several factors. For example, profit earned from product may change due to increase in cost of raw materials and competitive pricing. Sensitivity analysis in an LPP addresses the following issues:

1. Changes in the values of the RHS of a constraint.
2. Changes in the values of the objective function co-efficient.
3. Addition of a new variable.
4. Addition of new constraint.

The sensitivity analysis output for Example 15.1 using Excel Solver is shown in Table 15.4.

**TABLE 15.4** Sensitivity report from Excel Solver for Example 15.1

Microsoft Excel 14.0 Sensitivity Report						
Cells						
Cell	Name	Final Value	Reduced Cost	Objective Coefficient	Allowable Increase	Allowable Decrease
\$A\$2	X <sub>1</sub>	24	0	45	5	11.6666
\$B\$2	X <sub>2</sub>	16	0	50	17.5	5
Constraints						
Cell	Name	Final Value	Shadow Price	Constraint RHS	Allowable Increase	Allowable Decrease
\$A\$5	Corn Flour	20000	0.07	20000	4000	1000
\$A\$6	Sugar	28000	0	42000	1E + 30	14000
\$A\$7	Fruit and Nut	5200	0	10400	1E + 30	5200
\$A\$8	Ghee	9600	0.05	9600	400	1600
\$A\$9	DH	24	0	50	1E + 30	26
\$A\$10	TH	16	0	20	1E + 30	4

### 15.6.1 | Change in the RHS of a Constraint

Change in the RHS of a constraint is likely to change the feasible region of the LPP and may change the values of the objective function and the optimal solution. For example, consider the RHS of corn flour constraint; the current value of RHS (constraint RHS in Table 15.4) is 20,000 and the final value in the optimal solution is 20,000, that is, the entire quantity of corn flour is used; this means that the corn flour constraint is a binding constraint. Since the resource (corn flour) has been completely used, increase in this resource is likely to change the optimal solution and objective function value. Quantity by which the objective function value changes whenever the RHS of a constraint is changed by one unit is called the **shadow price**. In Table 15.4, the shadow price for corn flour is 0.07. This implies that for every one gram increase in the availability of corn flour, the profit will increase by INR 0.07. That is, 0.07 is the marginal value of one gram of corn flour. In the fruit and nut constraint only 5200 grams out of 10400 grams is used. That is, 5200 grams of fruit and nut is unused, so adding more fruit and nut will not add any value. Thus, the shadow price for the fruit and nut constraint is 0.


**IMPORTANT**

*Shadow price is the change in the objective function value for unit change in RHS of a constraint. It is the marginal value of the resource.*

Note that the shadow price will be non-zero only when the constraint is a binding constraint. In Example 15.1, the shadow price values for the binding constraints corn flour and ghee are 0.07 and 0.05, respectively, whereas the shadow price values for non-binding constraints sugar and fruit and nut are zero. In a binding constraint, the resource is completely used and thus additional resource is likely to change the optimal solution and the objective function value. On the other hand, in a non-binding constraint there is already excess resource (in case of less than or equal to constraint). Adding additional resource will not have any impact on the optimal solution and thus on the objective function value.


**IMPORTANT**

*Shadow price will be non-zero for a binding constraint and will be zero for a non-binding constraint.*

Note that the shadow price value is valid within an interval; as the RHS of a binding constraint is increased continuously, it is likely to become a non-binding constraint (excess resource). Similarly, when the RHS of a constraint is decreased continuously, the resource may become scarce and its marginal value is likely to increase. Thus, the current shadow price is valid only in an interval of the RHS value. In Table 15.4, for corn flour the allowable increase is 4000 and allowable decrease is 1000. That is, the shadow price value 0.07 is valid between a range of 19,000 (current RHS value – allowable decrease) and 24,000 (current value + allowable increase).

In Example 15.1, when the RHS is increased to 24001, the shadow price will become zero, whereas when the RHS is decreased below 19,000, the shadow price increases to 0.09 from 0.07. Now let us

consider a non-binding constraint such as sugar constraint. The current RHS is 42,000 and allowable increase is  $\infty$  and allowable decrease is 14,000. Since the allowable decrease is 14,000, the current shadow price value of 0 will change when the available sugar becomes less than 28,000 (current RHS – allowable decrease). The shadow price will become 0.035 from zero.




---

*Change in the RHS of a constraint will help us to understand how the shadow price changes. Shadow price remains same within a range of RHS values.*

### 15.6.2 | Impact of Change in the Coefficient Values in Objective Function

In the AH example, the objective function coefficients are the profits generated from two products DH and TH sold. The value of the objective function coefficient can change due to several factors such as increase in material cost, increase in production cost, etc. We would like to understand the impact of changes in the objective function coefficient values on the optimal solution. When the objective function coefficient value is changed, the optimal solution is likely to change. In Table 15.4, for the DH, the current value of coefficient is 45 and the allowable increase and decrease are 5 and 11.66, respectively. That is, as long as the coefficient (profit in this case) for DH lies between 33.34 (current value – allowable decrease) and 50 (current value + allowable increase), the current optimal solution ( $X_1 = 24, X_2 = 16$ ) will remain as the optimal solution. Similarly, for coefficient of  $X_2$  (coefficient of TH in objective function) the allowable increase and allowable decrease are 17.5 and 5, respectively. That is, for TH coefficient value between 45 and 67.5, the current optimal solution remains as the optimal solution. The range of objective function coefficient values for which the current optimal solution remains as optimal is called the *range of optimality*.

### 15.6.3 | 100% Rule

In Section 15.6.1, we discussed the impact of changes on RHS on the shadow price and thus on the optimal solution and objective function value. In Section 15.6.2, we discussed the impact of changes in the objective function coefficient value on the optimal solution. In both the cases, we assume that no other parameter values are changed. However, in real-life cases, parameter values may change simultaneously. If the simultaneous changes are made to the coefficients of non-basic variables within the range of allowable increase and decrease, then there will be no change to the current optimal solution. Similarly, if the simultaneous changes are made for non-binding constraint RHS values within the allowable increase and decrease, there will be no change in the basis (set of basic variables in the optimal solution). However, the simultaneous changes may be made to binding constraints and basic variables. In such cases we use the 100% rule described below (Bradley *et al.*, 1977; Wendell, 1985).

**100% rule for simultaneous changes in objective function coefficient values:** The following steps are used when more than one objective function coefficient values are changed.

1. Calculate  $R_i = \frac{\text{Actual change in the coefficient of variable } X_i}{\text{Allowable change for coefficient of variable } X_i \text{ in that direction}}$

If there is no change in  $X_p$ , then  $R_i = 0$ . Allowable change for coefficient of  $X_i$  in that direction implies that if coefficient of  $X_i$  is increased then we use the allowable increase. Similarly if the coefficient of  $X_i$  is decreased, then allowable decrease is used in calculating  $R_i$ .

2. Calculate  $R = \sum_{i=1}^n R_i$ .

If  $R \leq 1$ , then there will be no change in the current optimal solution. However, if  $R > 1$ , the test is inconclusive. That is, the current optimal solution may or may not change.

**100% rule for simultaneous changes in the right-hand side of constraint:** Changes in the right-hand side of a constraint is likely to change the current optimal basis and the shadow price values. The following steps are used when more than one right side of a constraint is changed. Let  $b_1, b_2, \dots, b_m$  be the current RHS value of constraints.

1. Calculate  $S_i = \frac{\text{Actual change in the RHS of Constraint } i}{\text{Allowable change for the constraint } i \text{ in that direction}}$

If there is no change in RHS then  $S_i = 0$ . Allowable change for RHS  $b_i$  (RHS of a constraint  $i$ ) in that direction implies that if  $b_i$  is increased then we use the allowable increase for constraint  $i$ . Similarly, if  $b_i$  is decreased, then allowable decrease is used in calculating  $S_i$ .

2. Calculate  $S = \sum_{i=1}^n S_i$ .

If  $S \leq 1$ , then there will be no change in current basis (and shadow price values). However, if  $S > 1$ , the test is inconclusive. That is, the basis may change.

#### 15.6.4 | Addition of a New Constraint

Adding a new constraint may change the feasible region of the problem and may change the current optimal solution. If the newly added constraint is redundant, then there will be no change in the current optimal solution.

#### 15.6.5 | Addition of a New Variable

Adding a new decision variable is likely to change many constraints of the problem and thus may change the optimal solution. The impact of adding a new constraint and new variable can be analysed using **complementary slackness theorem**.

### 15.7 | SOLVING A LINEAR PROGRAMMING PROBLEM USING GRAPHICAL METHOD

Whenever there are two decision variables in the model, an LPP can be solved graphically. An LPP with more than two variables can be solved when there are only two constraints by solving its dual problem. The following steps are used in graphical method:

1. Plot the feasible region using the constraints.
2. Draw the iso-profit (or iso-cost) line. Iso-profit (iso-cost) is the straight line represented by the objective function. If the objective function is maximization, then it is called iso-profit (iso implies equal, that is, the values on the line are equal) and in the case of minimization, the line formed by the objective function is called iso-cost line.
3. Move the iso-profit (iso-cost) line within the feasible region. The extreme point at which the iso-profit (iso-cost) leaves the feasible region is the optimal solution.

### EXAMPLE 15.2

#### Mumbai Perfumes

Mumbai Perfumes makes two types of perfumes called Chandan and Chamelee. Both perfume use two main ingredients: (a) Methyl benzoate and (b) Isobornyl cyclohexanol. Every 100 ml of Chandan requires 40 ml of Methyl benzoate and 80 ml of Isobornyl cyclohexanol. Every 100 ml of Chamelee requires 80 ml of Methyl benzoate and 20 ml of Isobornyl cyclohexanol. The total quantity of ingredients used is more than the volume of products produced due to production loss. The weekly availability of Methyl benzoate and Isobornyl cyclohexanol are 2000 ml and 3200 ml, respectively. The maximum production capacity of Mumbai Perfumes is 4000 ml per week. The revenue generated per 100 ml of Chandan is INR 2500 and 100 ml of Chamelee generates revenue of INR 2800. Formulate the problem and solve it using graphical method.

#### Solution:

Decision variables in this case are

$$\begin{aligned} X_1 &= \text{Amount of Chandan in ml to be produced} \\ X_2 &= \text{Amount of Chamelee in ml to be produced} \end{aligned}$$

The objective function for the problem is

$$\text{Maximize } 25X_1 + 28X_2$$

The constraints are

**Constraint for Methyl benzoate:** Every 100 ml of Chandan requires 40 ml of Methyl benzoate and every 100 ml of Chamelee requires 80 ml of Methyl benzoate. Weekly availability of Methyl benzoate is 2000 ml. The corresponding constraint is

$$0.4X_1 + 0.8X_2 \leq 2000$$

**Constraint for Isobornyl cyclohexanol:** Every 100 ml of Chandan requires 80 ml of Isobornyl cyclohexanol and every 100 ml of Chamelee requires 20 ml of Isobornyl cyclohexanol. Weekly availability of Isobornyl cyclohexanol is 3200 ml. The corresponding constraint is

$$0.8X_1 + 0.2X_2 \leq 3200$$

**Production constraint:** The maximum capacity is 4000 ml, that is

$$X_1 + X_2 \leq 4000$$

The linear programming formulation is

$$\text{Maximize } 25X_1 + 28X_2$$

subject to

$$0.4X_1 + 0.8X_2 \leq 2000$$

$$0.8X_1 + 0.2X_2 \leq 3200$$

$$X_1 + X_2 \leq 4000$$

$$X_1 \geq 0 \text{ and } X_2 \geq 0$$

### STEP 1 Identifying Feasible Region

Note that we can represent each constraint using a straight line. To draw a straight line, we need two points on the line. For example, for the Methyl benzoate constraint  $0.4X_1 + 0.8X_2 \leq 2000$ , the points (5000, 0) and (0, 2500) will be on the line  $0.4X_1 + 0.8X_2 = 2000$ . We can draw the line representing this constraint as shown in Figure 15.1. The area below this straight line satisfies this constraint. Similar logic can be used for Isobornyl cyclohexanol constraint and total production capacity constraint. In Figure 15.1, the feasible region is given by regions enclosed by vertices OABC.

### Iso-Profit line

Iso-profit line is the line represented by the objective function  $25X_1 + 28X_2$ . Assuming  $25X_1 + 28X_2 = K$  (where  $K$  is some constant), we can draw a straight line corresponding to the objective function (iso-profit line). In Figure 15.1, the iso-profit line is shown using a dotted line. Since the objective in this case is to maximize the revenue, we have to move the iso-profit line away from the origin. We can notice that the iso-profit line will leave the feasible region from extreme point (vertex) B. Extreme point B is an intersection between Methyl benzoate and total production constraint and at the point of intersection; both constraints will have equality sign. That is, at vertex B the following conditions exist:

$$0.4X_1 + 0.8X_2 = 2000$$

$$X_1 + X_2 = 4000$$

Solving above system of equations, we get  $X_1 = 3000$  and  $X_2 = 1000$ . That is, the optimal product mix is 3000 ml of Chandan and 1000 ml of Chamelee. The Excel Solver optimal solution output and sensitivity report are shown in Tables 15.5 and 15.6, respectively.

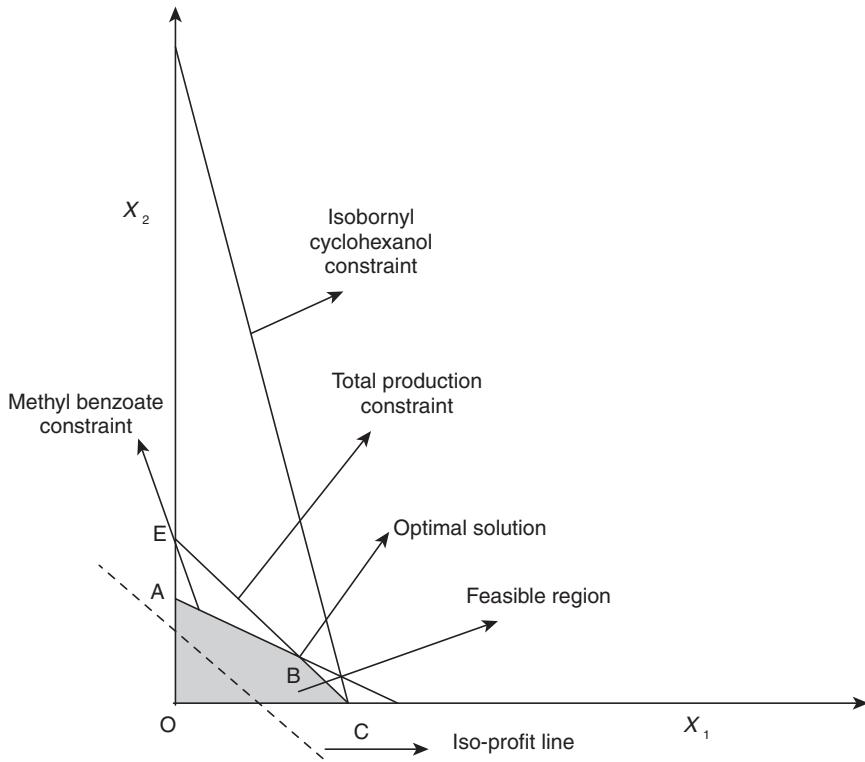


FIGURE 15.1 Graph describing feasible region, iso-profit line, and optimal solution.

**TABLE 15.5** Optimal solution using Excel Solver

$X_1$	$X_2$	Objective Function
3000	1000	103000
Constraints		
2000	2000	Methyl benzoate constraint
2600	3200	Isobornyl cyclohexanol constraint
4000	4000	Total production constraint

**TABLE 15.6** Sensitivity report**Microsoft Excel 14.0 Sensitivity Report**

Variable Cells						
Cell	Name	Final Value	Reduced Cost	Objective Coefficient	Allowable Increase	Allowable Decrease
\$A\$2	Chandan	3000	0	25	3	11
\$B\$2	Chamelee	1000	0	28	22	3

**TABLE 15.6** Sensitivity report—Continued

Cell	Name	Final Value	Shadow Price	Constraints		
				Constraint RHS	Allowable Increase	Allowable Decrease
\$A\$5	Constraints	2000	7.5	2000	1200	400
\$A\$6	Constraints	2600	0	3200	1E+30	600
\$A\$7	Constraints	4000	22	4000	428.5714	1500

## 15.8 | RANGE OF OPTIMALITY

Range of optimality is the range of objective function coefficient values for which the current optimal solution remains as optimal. In Table 15.6, the coefficient for Chandan is 25 and the allowable increase and decrease are 3 and 11, respectively. That is, for the coefficient value between 14 and 28, the current optimal solution ( $X_1 = 3000$  and  $X_2 = 1000$ ) will remain as optimal solution. Similarly, the allowable increase and decrease for coefficient of Chamelee are 22 and 3, that is, the current optimal solution will remain as optimal when the coefficient of Chamelee lies between 25 and 50.

Change in coefficient of a decision variable results in the change in slope of the objective function (iso-profit line). It is possible that as the slope of the iso-profit line changes, it may leave the feasible region from a different corner point rather than from the current corner point (vertex B) corresponding to the current optimal solution. In the current example, the optimal solution is given by vertex B. The objective function is

$$25X_1 + 28X_2$$

The slope of objective function is

$$X_1 = -(28/25)X_2$$

In general, for a LPP with two decision variables the slope of the objective function can be written as  $X_1 = (-B/A)X_2$ , where A and B are the coefficients corresponding to decision variables  $X_1$  and  $X_2$ , respectively. We are interested in finding the range of A and B for which the current optimal solution will remain as optimal solution. Note that, when the slope of the iso-profit line changes, it can leave the feasible region either from vertex A or vertex C which corresponds to slope of Methyl benzoate constraint and total production capacity constraint respectively. The slope of Methyl benzoate constraint is given by

$$X_1 = -(0.8/0.4)X_2 = -2X_2$$

The slope for production capacity constraint is given by

$$X_1 = -X_2$$

For the current optimal solution to remain as optimal, the slope of objective function should lie between the slopes of Methyl benzoate and total production capacity constraint. That is

$$-2 \leq - (B/A) \leq -1 \quad (15.1)$$

Equation (15.1) gives the range of optimality. It can be written as

$$1 \leq B/A \leq 2 \quad (15.2)$$

That is,

$$A \leq B \leq 2A$$

Since  $A = 25$ , the range of optimality of  $B$  is between 25 and 50, that is when the value of  $B$  lies between 25 and 50, the current optimal solution will remain as optimal solution. We can deduct this from the sensitivity output in Table 15.6. The current value of  $B$  is 28 and the allowable increase and decrease from Excel Solver output in Table 15.6 are 22 and 3, that is the range of  $B$  is 25 and 50.

The range of optimality for  $A$  is [from Eq. (15.2)]

$$1/2 \leq A/B \leq 1$$

Since  $B = 28$ , the range of  $A$  is between 14 and 28. In Table 15.6, the allowable increase and decrease are 3 and 11, respectively. Thus, the optimality range for  $A$  is between 14 and 28.

## 15.9 | RANGE OF SHADOW PRICE

Consider the constraint for Methyl benzoate. The shadow price for this constraint is 7.5 from Table 15.6. The allowable increase and decrease are 1200 and 400, respectively. Since the current RHS value for Methyl benzoate constraint is 2000, the range of the shadow price 7.5 is between 1600 ( $2000 - 400$ ) and 3200 ( $2000 + 1200$ ). This range can be derived from the graphical solution provided in Figure 15.1. When we increase the RHS of Methyl benzoate, the maximum increase possible is up to point  $E$ , beyond which the Methyl benzoate constraint will become redundant constraint. Point  $E$  corresponds to  $X_1 = 0$  and  $X_2 = 4000$ , which means the RHS value of Methyl benzoate constraint at vertex  $E$  is  $0.8 \times 4000 = 3200$  (which is 1200 more from current RHS value).

When we move the constraint towards origin, points  $B$  and  $C$  will merge and it will become the new optimal solution. Point  $C$  corresponds to  $X_1 = 4000$  and  $X_2 = 0$ . From Methyl benzoate constraint equation we get the RHS as  $0.4 \times 4000 = 1600$  (which is 400 less from the current RHS value).

## 15.10 | DUAL LINEAR PROGRAMMING

Dual linear programming is an important concept in LP which can be used to gain deeper insights about the problem under investigation. Any linear programming problem for the maximization of objective function can be expressed in the matrix form as

$$\text{Maximize } \mathbf{C}\mathbf{X} \quad (15.3)$$

$$\mathbf{A}\mathbf{X} \leq \mathbf{b} \quad (15.4)$$

$$\mathbf{X} \geq \mathbf{0} \quad (15.5)$$

where row vector  $\mathbf{C}$  is the coefficient of decision variables in the objective function,  $\mathbf{X}$  is the column vector representing the set of decision variables,  $\mathbf{A}$  is the matrix of coefficients in the constraints, and  $\mathbf{b}$  is the column vector representing the RHS of the constraints. We will call the problem defined in Eqs. (15.3)–(15.5) as the *primal problem*. For the Mumbai Perfumes, the primal problem formulation can be written in the matrix form as

$$\text{Maximize } \begin{bmatrix} 25 & 28 \end{bmatrix} \times \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \quad (15.6)$$

subject to constraints

$$\begin{bmatrix} 0.4 & 0.8 \\ 0.8 & 0.2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 2000 \\ 3200 \\ 4000 \end{bmatrix} \quad (15.7)$$

In the primal formulation, we try to allocate the resources (Methyl benzoate, Isobornyl cyclohexanol, and production capacity) between the products Chandan and Chamelee. That is, in the primal problem, we try to optimally allocate the resources between the products (resource allocation problem). Associated with a primal problem is a *dual problem* in which we try to find the marginal value of the resources (resource valuation problem, Naylor, 1966). Let  $Y_1$ ,  $Y_2$ , and  $Y_3$  be the marginal values of the resources Methyl benzoate, Isobornyl cyclohexanol, and production capacity (that is, value of facilities used in producing Chandan and Chamelee). That is

$$\begin{aligned} Y_1 &= \text{Marginal value of Methyl benzoate per ml} \\ Y_2 &= \text{Marginal value of Isobornyl cyclohexanol per ml} \\ Y_3 &= \text{Marginal value of production capacity per ml.} \end{aligned}$$

Consider a situation in which someone is interested in purchasing these resources from Mumbai Perfumes. The buyer will attempt to minimize the value she/he has to pay for these resources. Mumbai Perfumes has 2000 ml of Methyl benzoate, 3200 ml of Isobornyl cyclohexanol, and 4000 ml of production capacity every week. Thus, the objective of the dual problem (resource valuation problem) is given by

$$\text{Minimize } 2000Y_1 + 3200Y_2 + 4000Y_3, \quad (15.8)$$

However, Mumbai Perfumes will sell the resources, only under certain conditions. For example, using 0.4 ml of Methyl benzoate and 0.8 ml of Isobornyl cyclohexanol, and one unit of production capacity, 1 ml of Chandan is produced which generated a revenue (value) of INR 25. Since  $Y_1$ ,  $Y_2$ , and  $Y_3$  are the values of Methyl benzoate, Isobornyl cyclohexanol, and production capacity, the dual problem should satisfy the following constraint:

$$0.4Y_1 + 0.8Y_2 + Y_3 \geq 25 \quad (15.9)$$

The RHS in constraint Eq. (15.9) is the value generated by Mumbai Perfumes for every 1 ml of Chandan sold in the market. Similarly, using 0.8 ml of Methyl benzoate, 0.2 ml of Isobornyl cyclohexanol, and consuming one unit of production capacity, 1 ml of Chamelee is produced which generated a revenue (value) of INR 28. The corresponding constraint is

$$0.8Y_1 + 0.2Y_2 + Y_3 \geq 28 \quad (15.10)$$

The dual formulation (resource valuation formulation) is given by

$$\text{Minimize } 2000Y_1 + 3200Y_2 + 4000Y_3,$$

subject to the constraints

$$\begin{aligned} 0.4Y_1 + 0.8Y_2 + Y_3 &\geq 25 \\ 0.8Y_1 + 0.2Y_2 + Y_3 &\geq 28 \\ Y_1, Y_2 &\geq 0 \end{aligned}$$

The solution to the dual formulation using Excel Solver is shown in Table 15.7.

**TABLE 15.7** Optimal solution to the dual problem

$Y_1$	$Y_2$	$Y_3$	Objective Function
7.5	0	22	103000
<b>Constraints</b>			
25	25		Dual constraint corresponding to product Chandan
28	28		Dual constraint corresponding to product Chamelee

Note that the optimal values of  $Y_1$ ,  $Y_2$ , and  $Y_3$  are 7.5, 0, and 22, respectively, which are basically the values of shadow price for the constraints in the primal model (Table 15.6). The optimal value of the objective function in the dual problem is 103,000 which is same as the optimal value of the objective function in Table 15.6. The shadow price is also known as dual value (values of variables in dual problem).

### 15.10.1 | Conversion of a Primal Model to Dual Model

In this section, we will be explaining how a primal problem can be converted into a dual problem. The following steps are used in converting a primal maximization problem to a dual minimization problem:

1. Write the formulation in the standard form. For a maximization problem, in the standard form all the constraints should be less than or equal to constraints ( $\leq$ ). In the case of standard minimization, all the constraints should be greater than or equal to constraints ( $\geq$ ).
2. For each constraint in the primal problem, define a dual variable. Write the dual objective function  $\mathbf{Yb}$ . If the primal is maximization, the dual will be a minimization. Similarly, if the primal objective is minimization then the dual objective will be maximization.
3. For each variable in the primal identify the dual constraint  $\mathbf{YA} \geq \mathbf{C}$ .
4. Write the implicit constraint  $\mathbf{Y} \geq \mathbf{0}$ .

#### EXAMPLE 15.3

Convert the following primal to a dual formulation:

$$\begin{aligned} & \text{Maximize } X_1 - 2X_2 + 7X_3 \\ & \text{subject to constraints} \end{aligned}$$

$$\begin{aligned} & -X_1 + 2X_2 + 4X_3 \geq 18 \\ & X_1 + X_2 + X_3 = 10 \end{aligned}$$

#### Solution:

The first step in deriving dual formulation is to convert the primal to the standard form. Constraint 1 in primal is a greater than or equal to constraint, which can be converted to less than or equal to by multiplying it with  $-1$  on both sides, that is, the standard form of constraint 1 ( $-X_1 + 2X_2 + 4X_3 \geq 18$ ) is

$$X_1 - 2X_2 - 4X_3 \leq -18$$

The second constraint ( $X_1 + X_2 + X_3 = 10$ ) is an equality constraint, which can be split into two constraints as given below:

$$X_1 + X_2 + X_3 \leq 10$$

and  $X_1 + X_2 + X_3 \geq 10$  (that is,  $-X_1 - X_2 - X_3 \leq -10$ )

Thus the standard form of the primal problem is

$$\text{Maximize } X_1 - 2X_2 + 7X_3$$

subject to constraints

$$X_1 - 2X_2 - 4X_3 \leq -18$$

$$X_1 + X_2 + X_3 \leq 10$$

$$-X_1 - X_2 - X_3 \leq -10$$

The standard form of primal has 3 constraints and each constraint will have a dual variable. Let  $Y_1$ ,  $Y_2$ , and  $Y_3$  be the dual variables. The objective function for the dual is

$$\text{Minimize } -18Y_1 + 10Y_2 - 10Y_3$$

We now write the dual constraint for each primal variable. The dual constraint for primal variable  $X_1$  is

$$Y_1 + Y_2 - Y_3 \geq 1$$

Dual constraint for primal variable  $X_2$  is

$$-2Y_1 + Y_2 - Y_3 \geq -2$$

Dual constraint for primal variable  $X_3$  is

$$-4Y_1 + Y_2 - Y_3 \geq 7$$

Thus the dual formulation is

$$\text{Minimize } -18Y_1 + 10Y_2 - 10Y_3$$

$$Y_1 + Y_2 - Y_3 \geq 1$$

$$-2Y_1 + Y_2 - Y_3 \geq -2$$

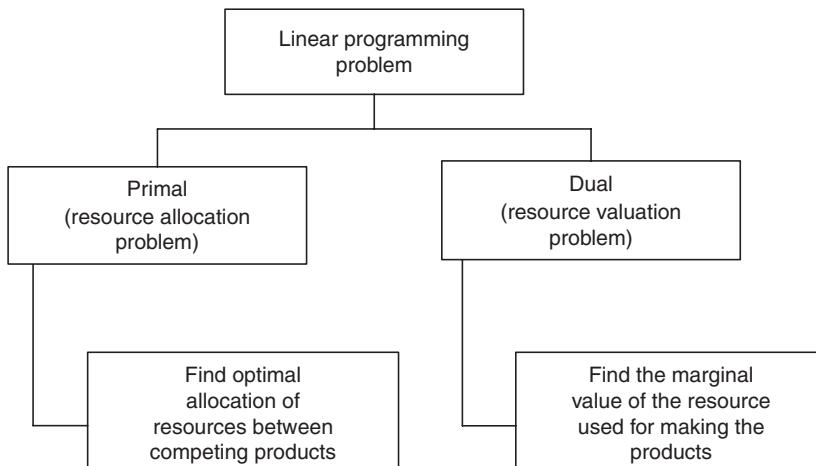
$$-4Y_1 + Y_2 - Y_3 \geq 7$$

$$Y_1, Y_2, \text{ and } Y_3 \geq 0$$

Note that in objective function as well as in all three constraints we have  $Y_2 - Y_3$ . This is due to the fact that these variables correspond to the equality constraint in the primal problem. We can replace  $Y_2 - Y_3$  in the dual model with a new variable  $Y_4$  ( $Y_4 = Y_2 - Y_3$ ). The variable  $Y_4$  will be an unrestricted variable (that is, a variable which take both positive and negative values). Whenever we have an equality constraint in the primal, the corresponding dual variable will be unrestricted. Similarly, if a decision variable is unrestricted in primal, the corresponding dual constraint will be an equality constraint.

## 15.11 | PRIMAL–DUAL RELATIONSHIPS

For every primal linear programming problem, we can formulate a corresponding dual problem. For easier understanding, assume that the primal is a resource allocation problem. Then the dual will be resource valuation problem as shown in Figure 15.2. For every variable in the primal, there will be dual constraint and for every constraint in the primal, there will be a dual variable.



**FIGURE 15.2** Primal–dual relationship.

### 15.11.1 | Weak Law of Duality

Consider the primal and dual formulation in its standard form as shown below:

Primal Formulation	Dual Formulation
Maximize $Z = CX$	Minimize $W = Yb$
$AX \leq b$	$YA \geq C$
$X \geq 0$	$Y \geq 0$

Constraint in the primal can be written as (multiplying with  $Y$  on both sides)

$$YAX \leq Yb \quad (15.11)$$

Constraint in the dual can be written as (multiplying with  $X$  on both sides)

$$YAX \geq CX \quad (15.12)$$

From Eqs. (15.11) and (15.12), we can write that

$$CX \leq Yb \quad (15.13)$$

The relationship in Eq. (15.13) is called weak law of duality. That is, if  $X$  is primal feasible and  $Y$  is dual feasible, then the objective function value of primal is less than or equal to the objective function value of dual.

**IMPORTANT**

As per the weak law of duality, if  $X(X_1, X_2, \dots, X_n)$  is a feasible solution in primal and  $Y(Y_1, Y_2, \dots, Y_m)$  is feasible solution in dual, then  $CX \leq Yb$ . That is, the objective function value of the primal will be less than or equal to the objective function value of the dual.

### 15.11.2 | Strong Law of Duality

According to the strong law of duality, if  $X^*$  is the optimal solution to the primal problem and  $Y^*$  is the optimal solution to the dual problem, then the following relationship is true:

$$CX^* = Y^*b$$

That is, the values of the objective function in primal and dual will be same when  $X^*$  is optimal solution for primal and  $Y^*$  is optimal solution for dual.

**IMPORTANT**

As per the strong law of duality, if  $X^*(X_1^*, X_2^*, \dots, X_n^*)$  is the optimal solution for primal problem and  $Y^*(Y_1^*, Y_2^*, \dots, Y_m^*)$  is the optimal solution for dual, then  $CX^* = Y^*b$ .

### 15.11.3 | Complementary Slackness Theorem

Complementary slackness theorem is useful for gaining insights about the LPP. Consider a primal linear programming problem with  $n$  decision variables and  $m$  constraints in the standard form. Then its dual will have  $m$  dual variables and  $n$  constraints. Complementary slackness theorem is stated as follows (Goldman and Tucker, 1956):

Let  $X = (X_1, X_2, \dots, X_n)$  be a feasible solution in primal and  $Y = (Y_1, Y_2, \dots, Y_m)$  be a feasible solution in dual. Also, assume that  $(S_1, S_2, \dots, S_m)$  denotes the slack variable in the primal for  $X = (X_1, X_2, \dots, X_n)$  and  $(W_1, W_2, \dots, W_n)$  denotes the surplus variables in dual. Then,  $X = (X_1, X_2, \dots, X_n)$  and  $Y = (Y_1, Y_2, \dots, Y_m)$  are optimal solutions if and only if  $X_j W_j = 0$  and  $Y_i S_i = 0$ .

That is, if a decision variable,  $X_j$ , is a basic variable (non-zero) in the primal optimal solution then the corresponding dual constraint will be a binding constraint (thus  $W_j = 0$ ). If the decision variable  $X_j$  is non-basic (that is,  $X_j = 0$ ), then the corresponding constraint will be non-binding (that is,  $W_j \neq 0$ ). Similarly, if a primal constraint is binding ( $S_i = 0$ ), then the corresponding dual variable  $Y_i$  will be non-zero. If a primal constraint is non-binding ( $S_i \neq 0$ ), then the corresponding dual variable,  $Y_i$ , will be zero.

We will be discussing the application of complementary slackness theorem using Example 15.4.

**IMPORTANT**

*Complementary slackness theorem is useful for understanding the concept of shadow price, reduced cost, impact of adding a new variable and a new constraint.*

**EXAMPLE 15.4****Application of Primal–Dual Relationship**

Thendral Krishnan (TK) is the co-founder and CEO of Great Western Vineyard (GWV), a vineyard established in Thekkady, Kerala. GWV cultivates three types of grapes, namely, (a) Malbec, (b) Cabernet, and (c) Riesling. The monthly production of Malbec, Cabernet, and Riesling at GWV is 350 kg, 300 kg, and 200 kg, respectively. Using these grapes, Thendral Krishnan makes three brands of wines, namely, (a) Chardonnay, (b) Blanc, and (c) Riesling W. The wines are made by mixing the three different types of grapes and two additives at various proportions. The amount of grapes and additives used to produce 1 litre of different brands of wines are shown in Table 15.8.

**TABLE 15.8** Quantity of grapes and additives used in 1 litre of wine

Wine	Grapes Amount in kg Used			Additive used in kg	
	Malbec	Cabernet	Riesling	Additive 1	Additive 2
Chardonnay	1.25	0.75	0.15	0.10	0.20
Blanc	0.80	1.80	0.25	0.05	0.15
Riesling W	0.4	0.6	2.0	0.08	0.12
Availability	350	300	200	120	80

The profit earned from wines Chardonnay, Blanc, and Riesling W are INR 400, 950, and 1350 per 1 litre, respectively. The following LP formulation is written to maximize the profit for GWV. Monthly maximum demand for Chardonnay, Blanc, and Riesling W is 1000, 750 and 250 litres, respectively.

$$X_1 = \text{Amount (in litres) of Chardonnay produced}$$

$$X_2 = \text{Amount (in litres) of Blanc produced}$$

$$X_3 = \text{Amount (in litres) of Riesling W produced}$$

The objective is

$$\text{Maximize } 400X_1 + 950X_2 + 1350X_3$$

subject to constraints

$$\begin{aligned}
 1.25 X_1 + 0.80 X_2 + 0.40 X_3 &\leq 350 \\
 0.75 X_1 + 1.80 X_2 + 0.60 X_3 &\leq 300 \\
 0.15 X_1 + 0.25 X_2 + 2.00 X_3 &\leq 200 \\
 0.10 X_1 + 0.05 X_2 + 0.08 X_3 &\leq 120 \\
 0.20 X_1 + 0.15 X_2 + 0.12 X_3 &\leq 80 \\
 X_1 &\leq 1000 \\
 X_2 &\leq 750 \\
 X_3 &\leq 250
 \end{aligned}$$

The Excel Solver output for the above formulation is given in Table 15.9. Note that Table 15.9 is a partial output with several missing values.

**TABLE 15.9** Microsoft Excel 14.0 Sensitivity Report

Variable Cells						
Cell	Name	Final Value	Reduced Cost	Objective Coefficient	Allowable Increase	Allowable Decrease
\$A\$2	$X_1$	0		400		1E +30
\$B\$2	$X_2$	139.1304	0	950	3100	50.26
\$C\$2	$X_3$	82.6087	0	1350	6250	859.09
Constraints						
Cell	Name	Final Value	Shadow Price	Constraint RHS	Allowable Increase	Allowable Decrease
\$A\$5	Malbec	144.34		350	1E +30	205.65
\$A\$6	Cabernet	300		300	473	240
\$A\$7	Riesling W	200		200	320.83	158.33
\$A\$8	Additive 1	13.56		120	1E +30	106.43
\$A\$9	Additive 2	30.78		80	1E +30	49.21
\$A\$10	Demand for $X_1$	0		1000	1E +30	
\$A\$11	Demand for $X_2$	139.13		750	1E +30	610.86
\$A\$12	Demand for $X_3$	82.60		250	1E +30	167.39

## Questions and Solutions

- (a) Thendral Krishnan can procure Malbec from Bangalore for INR 80 per kg. Should Thendral Krishnan buy Malbec? Justify your answer.

**Solution:**

From the sensitivity analysis we can see that the constraint corresponding to Malbec is non-binding. The availability is 350, but only 144.34 kg of Malbec is used in the optimal solution. Thus, GWV already has excess Malbec, so they should not buy additional Malbec from Bangalore.

- (b) What is the shadow price value for Cabernet? For what range of values this shadow price is valid?

**Solution:**

To calculate the shadow price, we have to formulate the dual and use the complementary slackness theorem. The primal has 8 constraints that means the dual will have 8 variables. Let  $Y_1, Y_2, \dots, Y_8$  be the dual variables corresponding to 8 primal constraints. The dual formulation is given by

Minimize  $350 Y_1 + 300 Y_2 + 200 Y_3 + 120 Y_4 + 80 Y_5 + 1000 Y_6 + 750 Y_7 + 250 Y_8$   
subject to constraints

$$1.25 Y_1 + 0.75 Y_2 + 0.15 Y_3 + 0.10 Y_4 + 0.20 Y_5 + Y_6 \geq 400$$

$$0.80 Y_1 + 1.80 Y_2 + 0.25 Y_3 + 0.05 Y_4 + 0.15 Y_5 + Y_7 \geq 950$$

$$0.40 Y_1 + 0.60 Y_2 + 2.00 Y_3 + 0.08 Y_4 + 0.12 Y_5 + Y_8 \geq 1350$$

$$Y_1, Y_2, \dots, Y_8 \geq 0$$

Note that in the optimal solution, only constraints 2 and 3 are binding constraints and all other constraints are non-binding constraint. That is, dual variables  $Y_2$  and  $Y_3$  are non-zero and all other dual variables are zero. Also, in the optimal solution  $X_1 = 0$  (non-basic variable), whereas  $X_2$  and  $X_3$  are basic variables (value greater than zero). Using complementary slackness theorem we can conclude that the constraint 1 is non-binding and constraints 2 and 3 are binding. Therefore, the above dual constraints can be written as

$$0.75 Y_2 + 0.15 Y_3 \geq 400$$

$$1.80 Y_2 + 0.25 Y_3 = 950$$

$$0.60 Y_2 + 2.00 Y_3 = 1350$$

Solving second and third equations, we get  $Y_2 = 452.8986$  and  $Y_3 = 539.1304$ . The shadow price for Cabernet is 452.8986.

- (c) A loyal customer of TK has asked for 20 litres of Chardonnay wine. What will be the impact on profit for GWV if TK decides to supply Chardonnay wine? TK would like to ensure that supplying Chardonnay will not impact the profitability of GWV. What will be your suggestion?

**Solution:**

In the current optimal solution, the value of  $X_1 = 0$  (that is quantity of Chardonnay produced is zero). Since,  $X_1$  is non-basic variable, it will have a non-zero reduced cost. Note that reduced cost is the value of the surplus variable in the corresponding dual constraint of  $X_1$ . The corresponding dual constraint is

$$0.75 Y_2 + 0.15 Y_3 \geq 400$$

Substituting  $Y_2 = 452.8986$  and  $Y_3 = 539.1304$ , we get the surplus variables as 20.5435. That is, the reduced cost is 20.5435. To ensure that GWV maintains its current profit, the profit from Chardonnay should be increased to 420.5435 per litre.

- (d) The supplier of grapes from Bangalore is ready to supply 500 kg Cabernet at the same price as the production cost of GWV resulting in no difference in profit. Should TK buy 500 kg from the Bangalore supplier? Justify your answer.

**Solution:**

In the current optimal solution, the entire 300 kg of Cabernet available for producing wines is used. Also, from question (b), we know that the shadow price for Cabernet constraint is 452.8986 and the allowable increase is 473. That is the shadow price of 452.8986 is valid for an additional 473 kg of Cabernet. So, GWV may buy 473 kg of Cabernet. Beyond that the shadow price is likely to decrease or even become zero.

- (e) The grape Malbec can be directly sold in the market at INR 70 per kg instead of using them in wine production. The monthly demand for Malbec grapes in Thekkady is 300 kg. What will be your suggestion to Thendral Krishnan?

**Solution:**

In the current optimal solution 144.34 kg is used out of 350 kg of Malbec that is available to GWV. Since 205.66 kg is not used in wine production, it may be sold in the market.

- (f) Due to Riesling crop failure in Nappa County, California USA, the demand for Riesling W increases significantly. TK decided to increase the profit on Riesling W to INR 2000. What will be impact of this change to the current optimal solution? What will be the impact on the profit?

**Solution:**

The current profit for Riesling W is INR 1350 and allowable increase from the sensitivity table is 6250. Since the proposed increase of INR 2000 is less than the allowable increase, there will be no change in the optimal solution. The profit will increase by  $82.60 \times 650$ .

- (g) Due to oversupply of wines in the market, Thendral Krishnan decides to decrease the profit on Blanc and Riesling W by INR 35 and INR 200, respectively. What will be impact of this reduction on the current optimal solution?

**Solution:**

Since more than one objective function coefficient value is changed simultaneously, we have to use the 100% rule. The allowable decrease in profit for Blanc and Riesling are 50.26 and 859.09, respectively.

$$R_1 = \frac{\text{Decrease in profit of Blanc}}{\text{Allowable decrease in profit of Blanc}} = \frac{35}{50.26} = 0.6964$$

$$R_2 = \frac{\text{Decrease in profit of Riesling W}}{\text{Allowable decrease in profit of Riesling W}} = \frac{200}{859.09} = 0.2328$$

$$R = R_1 + R_2 = 0.6964 + 0.2328 = 0.9292$$

Since  $R \leq 1$ , there will be no change in the optimal solution.

- (h) A new wine Merlot can be produced by mixing the various ingredients as shown below. The profit earned from Merlot is 990 per litre. Do you think GWV should produce this wine?

Malbec	Cabernet	Riesling	Additive 1	Additive 2
0.8	1.75	1.15	0.20	0.20

**Solution:**

Let  $X_4$  be the quantity of Merlot to be produced. Then the primal formulation can be written as

$$\text{Maximize } 400 X_1 + 950 X_2 + 1350 X_3 + 990 X_4$$

subject to constraints

$$1.25 X_1 + 0.80 X_2 + 0.40 X_3 + 0.8 X_4 \leq 350$$

$$0.75 X_1 + 1.80 X_2 + 0.60 X_3 + 1.75 X_4 \leq 300$$

$$0.15 X_1 + 0.25 X_2 + 2.00 X_3 + 1.15 X_4 \leq 200$$

$$0.10 X_1 + 0.05 X_2 + 0.08 X_3 + 0.20 X_4 \leq 120$$

$$0.20 X_1 + 0.15 X_2 + 0.12 X_3 + 0.20 X_4 \leq 80$$

$$X_1 \leq 1000$$

$$X_2 \leq 750$$

$$X_3 \leq 250$$

The dual constraint corresponding to primal variable  $X_4$  is

$$0.8Y_1 + 1.75Y_2 + 1.15Y_3 + 0.20Y_4 + 0.20Y_5 \geq 990$$

If the current solution remains as optimal solution (that is  $X_4 = 0$ ), then the above constraint will be feasible and non-binding. Substituting the values of dual variables in the aforementioned equation, we get

$$1.75 \times 452.8986 + 1.15 \times 539.1304 \geq 990$$

Since the new constraint is feasible and non-binding, we can conclude using complementary slackness theorem that  $X_4 = 0$ . That is, the current optimal solution remains as optimal.

## 15.12 | MULTI-PERIOD (STAGE) MODELS

In many real-life applications, the decisions are taken over several periods or stages. The decision maker has to take decisions during every stage of the planning horizon that optimizes the objective function. We will be explaining the multi-stage problem using Example 15.5 described below.

### EXAMPLE 15.5

Singham Puli is the Managing Director at Singham Puli & Sons (SPS), an investment advisory company. One of the customers has INR 1,000,000 to invest and has approached SPS for advice on optimal investment plan that will maximize the return after 5 years. The following assets are available for investment over next five years.

- (a) Option A is available at the beginning of every year and every 1 rupee invested in option A will return 1.2 at the end of 2 years after maturity.
- (b) Option B is available at the beginning of years 1, 2, and 3 and every 1 rupee invested in option B will return 1.5 at the end of 3 years after maturity.
- (c) Option C is available at the beginning of years 3, 4, and 5 and every 1 rupee invested in option C will return 1.10 at the end of 1 year.

Design a LP formulation that Singham Puli can use to maximize the money at the beginning of year 6.

#### **Solution:**

This problem is an example of multi-stage LPP. Multi-stage models have the following characteristics:

1. An initial condition which serves as the input to the first stage of the problem. In this example, the customer has INR 1,000,000 to invest which is the initial condition.

2. Decisions to be made during each stage of the problem. Decision variables in this example are the amount of money to be invested in options A, B, and C (if available) every year.
3. Link between different stages. Link in this problem is the returns that arrive at the beginning of each year from previous investments.

The model formulation is as follows: Let

- $A_t$  be amount of money invested in option A at the beginning of year  $t$  ( $t = 1, 2, 3, 4$ ).
- $B_t$  be amount of money invested in option B at the beginning of year  $t$  ( $t = 1, 2, 3$ ).
- $C_t$  be amount of money invested in option C at the beginning of year  $t$  ( $t = 3, 4, 5$ ).
- $N_t$  be amount not invested in any option at the beginning of year  $t$  ( $t = 1, 2, 3$ ).

The variable  $N_t$  is used here since it is possible that the decision maker may not invest in the current period in anticipation of a better investment option in the future.

#### **STAGE 1** (Beginning Year 1)

---

At the beginning of year 1 the available money is 1,000,000. The decision maker has to decide how much to invest in options A and B (he may also decide not to invest with an anticipation of better option in the future). The corresponding constraint is

$$A_1 + B_1 + N_1 = 1,000,000$$


---

#### **STAGE 2** (Beginning Year 2)

---

At the beginning of year 2 the available fund will be  $N_1$  and the available options are  $A_2$ ,  $B_2$ , and  $N_2$ .

$$A_2 + B_2 + N_2 = N_1$$


---

#### **STAGE 3** (Beginning Year 3)

---

At the beginning of year 3, the amount invested in option A in year 1 ( $A_1$ ) will be available and it will be  $1.2A_1$ . The constraint for Stage 3 is

$$A_3 + B_3 + C_3 = N_2 + 1.2 A_1$$

Note that we have not used  $N_3$  in stage 3 constraint since all three investment options are available at the beginning of stage 3.

---

#### **STAGE 4** (Beginning Year 4)

---

The constraint for year 4 is

$$A_4 + C_4 = 1.2 A_2 + 1.5 B_1 + 1.10 C_3$$


---

**STAGE 5** (*Beginning Year 5*)

The constraint for year 5 is

$$C_5 = 1.2A_3 + 1.5B_2 + 1.10C_4$$

The objective is to maximize the money that will be available at the beginning of year 6, that is

$$\text{Maximize } 1.2A_4 + 1.5B_3 + 1.10C_5$$

The optimal solution to the problem is shown in Table 15.10.

**TABLE 15.10** Optimal solution using Excel Solver

$A_1$	$A_2$	$A_3$	$A_4$	$B_1$	$B_2$	$B_3$
0	0	0	0	1000000	0	0
$C_3$	$C_4$	$C_5$	$N_1$	$N_2$	Objective	
0	1500000	1650000	0	0	1815000	
<b>Constraints</b>						
1000000	1000000	Year 1				
0	0	Year 2				
0	0	Year 3				
1500000	1500000	Year 4				
1650000	1650000	Year 5				

The optimal solution is

$$\begin{aligned} B_1 &= 1000000 \\ C_4 &= 1500000 \\ C_5 &= 1650000 \end{aligned}$$

and the maximum return that can be obtained through this investment is 1815000.

### 15.13 | LINEAR INTEGER PROGRAMMING (ILP)

Integer linear programming problems are linear programming problems with additional constraint that one or more decision variable can take only integer values. ILP is an important class of problem in prescriptive analytics since many industrial problems are integer programming problems. Integer programming problems are categorized under the following three categories:

- Pure integer programming:** In pure integer programming, all the decision variables can take only integer values.
- Mixed integer programming:** In a mixed integer programming, not all decision variables are integers; few of them can take non-integer values.

3. **Zero-one programming:** In a zero-one (or binary) programming problem, the decision variable can take either zero or one.

There are many algorithms used for solving integer programming problems and the simplest is called LP relaxation approach in which the integer constraints are ignored and the problem is solved as regular linear programming problem using simplex and other algorithms. Any decision variable that has a non-integer value obtained by solving relaxed problem is rounded to nearest integer that is feasible. However, LP relaxation will not guarantee optimal solution to the integer programming problem. One of the most popular algorithms used for solving integer programming is branch and bound (B and B) algorithm. Branch and Bound algorithm is a divide-and-conquer problem-solving strategy which is explained in the following section.

### 15.13.1 | Branch and Bound Algorithm

Branch and Bound algorithm starts with a relaxed problem of the integer programming (IP) problem to be solved. In general, relaxed problems have the following properties (Lawler and Wood, 1966):

1. Let  $P$  be the original integer programming problem with feasible region  $F$ .
2. The relaxed problem,  $P_R$ , is a problem which is easier to solve with feasible region  $F_R$  such that  $F \subset F_R$ .
3. For an ILP with maximization objective, the optimal solution of relaxed problem  $P_R$  is an upper bound for the original problem  $P$ . That is, the optimal solution of the IP problem cannot exceed the optimal solution of the relaxed problem.
4. For an ILP with minimization objective, the optimal solution for relaxed problem  $P_R$  is a lower bound for the original problem  $P$ . That is, the optimal solution of the IP problem cannot be less than that of the relaxed problem.

Branch and Bound uses the following steps to solve an ILP (Lawler and Wood, 1966):

1. **Identify the appropriate relaxed problem:** For example, if the problem is pure integer programming problem, then we ignore the integer constraints. If the problem is zero-one programming problem, then we replace the binary constraint for each decision  $X_i$  with  $0 \leq X_i \leq 1$  in the relaxed problem. In case of mixed integer programming problem, relaxation is used only for integer variables.
2. **Branching step:** Divide the problem into two sub-problems by introducing new constraints. For example, let decision variable  $X_k$  be non-integer in the optimal solution of the relaxed problem. Then we create two sub-problems by introducing the following constraints:

$$X_k \leq \lfloor X_k \rfloor \text{ and } X_k \geq \lceil X_k \rceil$$

where  $\lfloor X_k \rfloor$  is the nearest integer smaller than  $X_k$  and  $\lceil X_k \rceil$  is the nearest integer greater than  $X_k$ . Let  $X_k = 4.2$ . Then  $\lfloor X_k \rfloor = 4$  and  $\lceil X_k \rceil = 5$ .

3. **Bounding step:** For each sub-problem created in branching step we calculate the upper and lower bounds (when available) and fathom (discontinue from further branching) a sub-problem whenever possible. In case of maximization, lower bound is a candidate solution, that is, it satisfies all the integer constraints of the original problem.
4. **Stopping criteria:** If all branches are fathomed, then the current best solution is the optimal solution.

Fathoming of a branch is done using the following conditions:

1. The sub-problem created by additional constraint is infeasible.
2. The solution to the sub-problem satisfies the integer constraints of the original problem (candidate solution).
3. The optimal solution to current sub-problem is less than (greater than) the best candidate solution obtained so far for maximization (minimization) problem.

#### EXAMPLE 15.6

A manufacturer produces two types of kitchen closets (model 1 and 2) using two processes: (a) Assembly and (b) Finishing. Model 1 requires 18 man hours in assembly and 14 man hours in finishing. Model 2 requires 14 man hours in assembly and 40 man hours in finishing. Every week 112 man hours in assembly and 140 man hours of finishing are available. The profit generated from model 1 is INR 400 and that from model 2 is INR 900. Find the optimal integer solution that will maximize the profit using branch and bound algorithm.

#### Solution:

Let  $X_1$  and  $X_2$  be the number of model 1 and 2 produced by the manufacturer. The integer programming formulation is given by

$$\text{Maximize } Z = 400X_1 + 900X_2$$

subject to constraints

$$\begin{aligned} 18X_1 + 14X_2 &\leq 112 && \text{(assembly constraint)} \\ 14X_1 + 40X_2 &\leq 140 && \text{(finishing constraint)} \end{aligned}$$

$X_1, X_2$  are integers and  $\geq 0$ .

**Relaxed Problem:** The corresponding relaxed problem (say problem 1) is given by

$$\text{Maximize } Z = 400X_1 + 900X_2$$

subject to constraints

$$\begin{aligned} 18X_1 + 14X_2 &\leq 112 \\ 14X_1 + 40X_2 &\leq 140 \end{aligned}$$

$X_1, X_2 \geq 0$  (integer conditions are relaxed)

Solving the relaxed problem, we get the optimal  $Z = 3558.779$ ,  $X_1 = 4.81$ , and  $X_2 = 1.82$  (rounded to two decimals).

**Branching Step:** Since both  $X_1$  ( $= 4.81$ ) and  $X_2$  ( $= 1.82$ ) are non-integers, we can choose any one of them to create sub-problems by introducing additional constraints.

### Problem 2

Maximize  $Z = 400X_1 + 900X_2$   
subject to constraints

$$18X_1 + 14X_2 \leq 112$$

$$14X_1 + 40X_2 \leq 140$$

$$X_1 \leq 4$$

$$X_1, X_2 \geq 0$$

Optimal solution to problem 2 is  $Z = 3490$ ,  $X_1 = 4.0$ , and  $X_2 = 2.1$ .

### Problem 3

Maximize  $Z = 400X_1 + 900X_2$   
subject to constraints

$$18X_1 + 14X_2 \leq 112$$

$$14X_1 + 40X_2 \leq 140$$

$$X_1 \geq 5$$

$$X_1, X_2 \geq 0$$

Optimal solution to problem 3 is  $Z = 3414.28$ ,  $X_1 = 5.0$ , and  $X_2 = 1.57$ .

Figure 15.3 shows the branching of the relaxed problem (problem 1) into sub-problems 2 and 3.

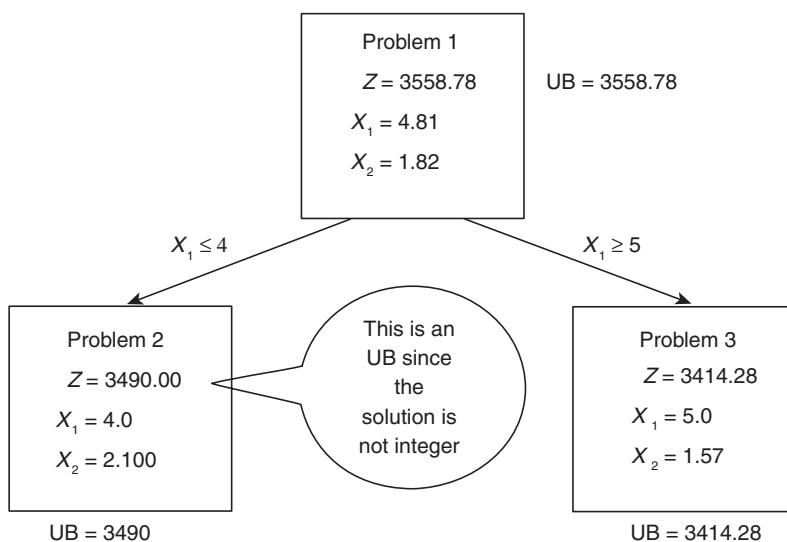


FIGURE 15.3 Branching step in B and B algorithm.

The optimal solution to the relaxed problem (problem 1) is 3558.78. This is an upper bound (UB) since any sub-problem created from this problem will have an optimal solution less than or equal to 3558.78. Problem 2 is created by using additional constraint  $X_1 \leq 4$  and its UB is 3490. Sub-problem 3 is created by using constraint  $X_1 \geq 5$  and its UB is 3414.28.

Note that after the first branching, we still do not have a candidate solution (solution in which all the decision variables are integers). So, we continue to branch, and the final branch and bound tree are shown in Figure 15.4.

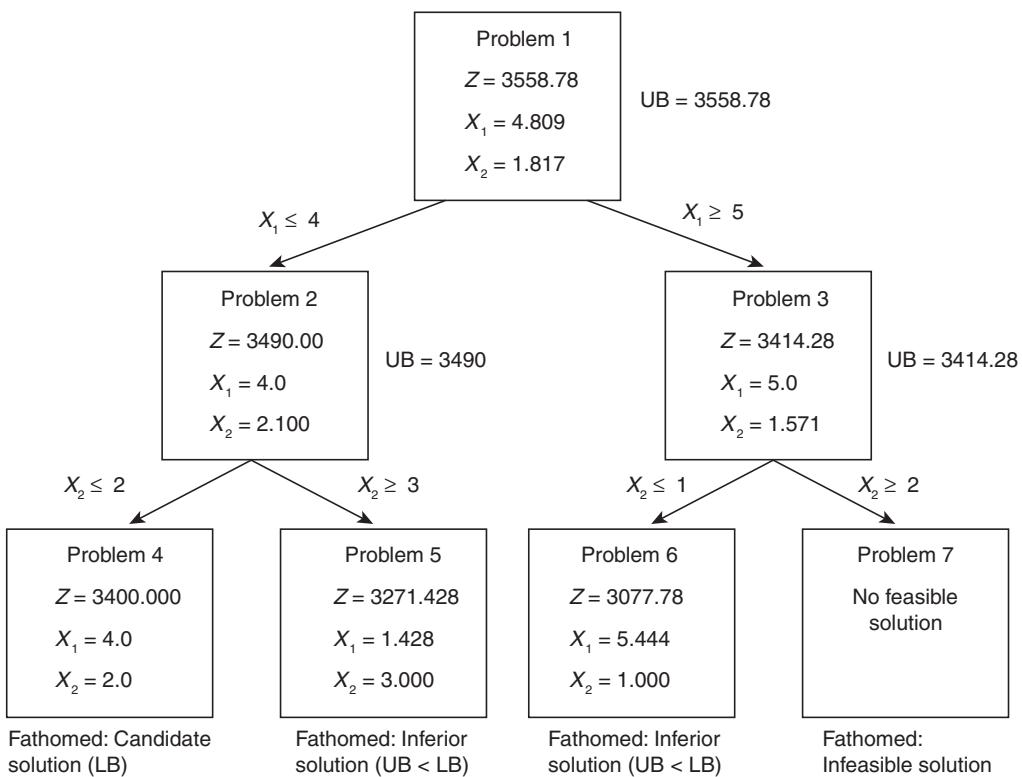


FIGURE 15.4 Branch and bound tree for the Example 15.6.

**Bounding Step (Tree Pruning):** In the branch and bound tree shown in Figure 15.4 consider the sub-problem 4. The additional constraint in this case is  $X_2 \leq 2$  (in addition to constraints defined in sub-problem 2). The optimal solution is  $X_1 = 4$ ,  $X_2 = 2$ , and  $Z = 3400$ . Since both variables are integers, this is a candidate solution. This node will be fathomed since any branching will result in objective function values less than or equal to 3400. Since 3400 is a candidate solution, it is also the lower bound (LB) for the original ILP problem. That is, the optimal solution to the original ILP cannot be less than 3400.

In sub-problem 5, the optimal  $Z$  (optimal solution to relaxed problem) value is 3271.428 and  $X_1 = 1.428$  and  $X_2 = 3$ . Note that 3271.428 is an upper bound for the sub-problem 5, any further branching will result in reduced values for the objective function value. Since this value (UB for sub-problem 5) is less than the known candidate solution (LB, solution in sub-problem 4), we fathom this branch. Similarly, sub-problem 6 is also fathomed since value of  $Z$  is less than the candidate solution and sub-problem 7 is infeasible. Thus, we have fathomed all branches. The best candidate solution is  $X_1 = 4$ ,  $X_2 = 2$ , and  $Z = 3400$ , which is the optimal solution to the problem.

### 15.13.2 | Branching Strategies in Branch and Bound Algorithm

One of the decisions to be taken while using branch and bound algorithms is selection of a node for further branching among several nodes that are not fathomed. The following strategies are used for selection of a node for branching:

- 1. Depth First Search (DFS):** In this strategy the branching is carried out on the most recently branched node (sub-problem) as far as possible (that is, till a sub-problem is fathomed). Once it reaches dead end (all problems in that branch are fathomed), it backtracks to a unfathomed branch and continues with depth first search. DFS is also known as backtracking or last-in-first-out (LIFO) strategy. DFS helps to reach a candidate solution as soon as possible, which will help to fathom sub-problems from other branches.
- 2. Breadth First Search (BFS):** Breadth first strategy starts from root node and explores first level of branches before moving to second level of nodes. That is, BFS uses first-in-first-out (FIFO) strategy to explore the tree. BFS is also known as jump tracking since the algorithm jumps to best known solution to the relaxed problem for further branching.

## 15.14 | MULTI-CRITERIA DECISION-MAKING (MCDM) PROBLEMS

In many practical problems, one may like to optimize more than one objective function simultaneously. For example, a company may like to design a highly reliable product at the lowest cost. However, maximizing reliability of the product may involve use of better material and thus may increase the cost which we would like to decrease. That is, in MCDM problem the objectives are likely to be conflicting with one another. There are many techniques used for solving MCDM such as analytic hierarchy process (AHP) and goal programming. In this chapter, we will be discussing goal programming technique for solving MCDM problems.

### 15.14.1 | Goal Programming

Goal programming is used when the problem has more than one objective function. Each objective is set as goals with targets which serve as upper and/or lower bounds. The following steps are used in building a goal programming model (Charnes *et al.*, 1955):

- Identify the objectives of the problems and the corresponding target. Each objective is considered as a goal and penalty is assigned for violation of each goal.
- Identify decision variables of the problem.
- For each goal (objective) write the goal constraint using deviation variables. Goal constraints are soft constraints that can be violated. Deviation variables capture the deviation from the target set for each goal.
- Write system constraints which are hard constraints that cannot be violated.
- Write the objective function as a minimization of weighted sum deviation variables, where the weights are the penalties assigned for violation of goals.

### *Preemptive and Non-Preemptive Goal Programming*

Goal programming problems are classified into preemptive and non-preemptive goal programming. In preemptive goal programming, the goals are defined as hierarchy of goals. That is goals are defined at different priority levels. On the other hand, in a non-preemptive goal programming, all goals are treated as equally important.

#### **EXAMPLE 15.7**

Miss Moneypenny is in charge of a retail store that sells movie merchandise in a store located in Bond Street, London. She has to allocate shelf space (in cubic feet) for different products that she would like to sell in her retail store based on (a) shelf space required, (b) expected profit per week, and (c) attractiveness index of the product (in a scale of 1–10, higher value indicates higher attractiveness). To ensure product visibility if a product is chosen as part of the assortment, certain minimum space should be allocated. The details are given in Table 15.11.

**TABLE 15.11** Product information consisting of space required, profit, and attractiveness index

Product	1	2	3	4	5	6
Space needed in cubic feet	240	240	160	200	320	120
Expected profit per week in British Pounds	1800	1600	1500	1450	1380	1420
Attractiveness index	7	6	6	7	8	7

She has set the following three goals:

- The profit earned should be at least 8000 British Pounds.
- The total attractiveness index should be at least 35.
- The total space cannot exceed 1000 cubic feet.

All goals are treated equally important. Miss Moneypenny would like to sell a maximum of 4 products. Penalty for missing every one dollar on profit is 1 unit, penalty

for missing one unit of attractiveness index is 5 units, and penalty for exceeding 1 cubic feet of shelf space is 10 units. Formulate a goal programming model that Miss Moneypenny can use to meet all the constraints and goals.

**Solution:**

In this case we have three goal constraints (soft constraints) and one system constraint (maximum products to be sold). The decision variables are:

$$Y_i = \begin{cases} 1 & \text{if product } i \text{ is chosen as part of the assortment} \\ 0 & \text{otherwise} \end{cases}$$

**Deviation variables and goal constraint:** Consider the goal that the profit earned should be at least 8000 British Pounds (BP); 8000 BP is the target and is a lower bound. Deviation variables are variables used for measuring the difference between the achieved value (profit in this case) and the target value (8000 pounds in this case). Let  $D_1$  be the deviation variable for the profit goal. It can be written as

$$D_1 = 1800Y_1 + 1600Y_2 + 1500Y_3 + 1450Y_4 + 1380Y_5 + 1420Y_6 - 8000$$

Note that,  $D_1$  can take both positive and negative values (unrestricted variable); however, to apply simplex algorithm all variables should be non-negative. An unrestricted variable can be incorporated in the model by using two non-negative artificial variables. That is, the unrestricted variable  $D_1$  can be written as

$$D_1 = D_1^+ - D_1^-$$

where  $D_1^+$  and  $D_1^-$  are non-negative values. When  $D_1$  is positive,  $D_1^+$  will be positive and  $D_1^-$  will be zero. Similarly, when  $D_1$  is negative,  $D_1^+$  will be zero and  $D_1^-$  will be positive. The profit goal can be written as

$$1800Y_1 + 1600Y_2 + 1500Y_3 + 1450Y_4 + 1380Y_5 + 1420Y_6 - D_1^+ + D_1^- = 8000$$

Note that  $D_1^+ > 0$  implies that the achieved profit is more than 8000 whereas  $D_1^- > 0$  implies that the achieved profit is less than 8000 and thus we would like to minimize  $D_1^-$ .

**Attractiveness index goal:** The minimum attractiveness should be at least 35. The corresponding goal constraint is

$$7Y_1 + 6Y_2 + 6Y_3 + 7Y_4 + 8Y_5 + 7Y_6 - D_2^+ + D_2^- = 35$$

where  $D_2 = D_2^+ - D_2^-$  is the deviation variable for the attractiveness index. Since we would like the attractiveness index to be at least 35,  $D_2^- > 0$  is undesirable and we would like to minimize.

**The total space goal:** The total space cannot exceed 1000 cubic feet. So

$$240Y_1 + 240Y_2 + 160Y_3 + 200Y_4 + 320Y_5 + 120Y_6 - D_3^+ + D_3^- = 1000$$

where  $D_3 = D_3^+ - D_3^-$  is the deviation variable for space goal. In this case,  $D_3^+ > 0$  implies using more than 1000 cubic feet space and thus we would like to minimize  $D_3^+$ .

### The system constraint:

$$Y_1 + Y_2 + Y_3 + Y_4 + Y_5 + Y_6 \leq 4$$

The objective is

$$\text{Minimize } D_1^- + 5 \times D_2^- + 10 \times D_3^+$$

The optimal solution obtained using Excel Solver is shown in Table 15.12.

**TABLE 15.12** Optimal solution to goal programming Example 15.7

$Y_1$	$Y_2$	$Y_3$	$Y_4$	$Y_5$	$Y_6$	
1	1	1	1	0	0	
$D_1^+$	$D_1^-$	$D_2^+$	$D_2^-$	$D_3^+$	$D_3^-$	OBJ
0	1650	0	9	0	160	1695
<b>Constraints</b>						
8000	8000			Profit Goal		
35	35			Attractiveness Index Goal		
1000	1000			Total Space Goal		
4	4					

Total penalty in the optimal solution is 1695. The value of  $D_1^- = 1650$ , that is the achieved profit is  $8000 - 1650 = 6350$ . The value of  $D_2^- = 9$ , that is the achieved attractiveness index is  $35 - 9 = 26$ .  $D_3^- = 160$ , which implies the total space required is 840.

### EXAMPLE 15.8

#### Investment Decision

One of the customers of ‘Share Khan and Sons (SKS)’, an investment advisory company, has approached them for advisory services. The customer would like to invest INR 1,00,000 for a period of one year and would like to achieve a target return of INR 16,000 at the end of one year. The customer also would like to restrict the risk to a maximum of 500 units. SKS has identified shares of four companies that the customer may invest, their share price, return in rupees per share, and risk (Table 15.13). Between risk and return goals, the customer would like to give top priority to the risk goal. Formulate a pre-emptive goal programming model to solve the problem.

**TABLE 15.13** Share price, return, and risk information

Company	Price Per Share	Annual Return	Risk
Company 1	25	3.5	0.20
Company 2	18	2.2	0.10
Company 3	19	2.75	0.08
Company 4	33	4.1	0.15

**Solution:**

Let

$$X_i = \text{Number of shares of company } i \text{ to be purchased}$$

Return goal constraint can be written as

$$3.5X_1 + 2.2X_2 + 2.75X_3 + 4.1X_4 - D_1^+ + D_1^- = 16000$$

$D_1^- \geq 0$  implies that the return is less than 16000, so we would like to minimize  $D_1^-$ .

Risk goal constraint can be written as

$$0.2X_1 + 0.1X_2 + 0.08X_3 + 0.15X_4 - D_2^+ + D_2^- = 500$$

$D_2^+ \geq 0$  implies that the risk is greater than 500, so we would like to minimize  $D_2^+$ .

System constraint (money available for investment) is

$$25X_1 + 18X_2 + 19X_3 + 33X_4 \leq 100000$$

Let  $P_1$  and  $P_2$  be the penalty for violation of return and risk goals. Since risk goal is the top priority goal, we will choose  $P_2 >> P_1$ .

The objective function value is

$$\text{Minimize } P_1 D_1^- + P_2 D_2^+$$

Optimal solution using Excel Solver is shown in Table 15.14 ( $P_1 = 1$ ,  $P_2 = 100$ ).**TABLE 15.14** Optimal goal programming solution for Example 15.8

$X_1$	$X_2$	$X_3$	$X_4$	
0	0	5263	0	
$D_1^+$	$D_1^-$	$D_2^+$	$D_2^-$	Objective
0	1526.75	0	78.96	1526.75
<b>Constraints</b>				
16000	16000			Return Goal
500	500			Risk Goal
99997	100000			Available money for investment

The optimal solution is  $X_1 = X_2 = X_4 = 0$  and  $X_3 = 5263$ . The total penalty is 1526.75, that is the return is 14473.25 and the risk is  $500 - 78.96 = 421.04$ .

**SUMMARY**

1. Prescriptive analytics is used to find best solution/decision to a problem. Prescriptive analytics is the frontier of analytics capability.
2. Operations research techniques such as linear programming, integer programming, and goal programming are frequently used for solving prescriptive analytics problems.
3. Many prescriptive analytics problems such as travelling salesman problem, bin-packing problem, facility location problem do not have efficient algorithms to solve as the size of problem increases and are big data problems within prescriptive analytics.
4. Linear programming is one of the most popular prescriptive analytics technique used for solving several analytics problems such as product mix, transportation, revenue management, assortment planning, shelf space allocation, etc.
5. The concept of shadow price provides insights about the marginal value of a resource that will be very useful in decision making.
6. Integer programming is another prescriptive analytics problem solving technique that can be used for several organizational problems such as capital budgeting, facility location. Integer programming problems are usually solved using branch and bound algorithms.
7. Many prescriptive analytics problems may have multiple objectives. Problems with multiple objectives are solved using goal programming.

**MULTIPLE CHOICE QUESTIONS**

1. In a linear programming problem
 

(a) Feasible region is always bounded	(b) Feasible region is unbounded
(c) Feasible region is a convex set	(d) Feasible region is a concave set
2. The value of non-basic decision variable in the optimal solution is
 

(a) Greater than zero	(b) Zero
(c) Greater than or equal to zero	(d) Less than or equal to zero
3. Shadow price of a constraint is non-zero when
 

(a) Constraint is non-binding	(b) Slack variable is non-zero
(c) Constraint is redundant	(d) Constraint is binding
4. Reduced cost is
 

(a) Zero for basic variable and non-zero for non-basic variable	(b) Non-zero for basic variable and zero for non-basic variable
(c) Zero for both basic and non-basic variable	(d) Non-zero for both basic and non-basic variable
5. If the decision variable has non-zero value in the primal optimal solution, then
 

(a) The corresponding dual constraint will be a binding constraint	(b) The corresponding dual constraint will be a non-binding constraint
(c) The corresponding constraint dual variable is non-zero	(d) The corresponding dual constraint may or may not be binding
6. The value of the dual variable is
 

(a) Reduced cost corresponding to the primal variable	(b) Shadow price corresponding to the primal constraint
(c) Slack variable in the primal constraint	(d) Surplus variable in the primal constraint

7. If the primal constraint is equality constraint then
  - (a) The dual variable will be positive
  - (b) The dual variable will be negative
  - (c) The dual variable will be unrestricted
  - (d) The dual variable will be zero
8. When the value of RHS of a non-binding constraint is changed
  - (a) The shadow price will not change within the range of allowable increase and decrease
  - (b) The shadow price will be zero change within the range of allowable increase and decrease
  - (c) The shadow price will not change.
  - (d) The shadow price will be zero.

### EXERCISES

1. Perumal Nadar & Sons (PNS) is a manufacturer of Ayurvedic supplements prepared as per the ancient books on Ayurveda. One of their main product is Chyawanprash prepared using sugar, ghee, honey, amla (Indian Gooseberry), and herbs. Few additional ingredients are added based on variants of Chyawanprash. For example, dried fruits and nuts are added in the fruit and nut flavour and chocolate is added in case of chocolate Chyawanprash. PNS is interested in finding the optimal product mix for three variants of Chyawanprash: (a) Regular, (b) Fruit and Nut, and (c) Chocolate. Ingredients such as sugar and ghee are readily available in the market; however, supply is constrained for ingredients such as honey, amla, and herbs. The quantity of ingredients required for different types of Chyawanprash is shown in Table 15.15 per 1 kg of Chyawanprash. The profit earned from regular, fruit and nut, and chocolate flavour per kg is 85, 102, and 105, respectively. The monthly minimum demand for regular, fruit and nut, and chocolate is 100 kg, 100 kg, and 200 kg, respectively.

**TABLE 15.15** Data related to different Chyawanprash produced by PNC

	Chyawanprash			Availability
	Regular	Fruit and Nut	Chocolate	
Honey (in grams)	70	75	80	15,000
Amla (in grams)	140	165	170	50,000
Herbs (in grams)	120	120	145	20,000

- (a) Develop an LP model to maximize the profit and solve it using Excel Solver.
- (b) What is the shadow price for the constraints honey, amla, and herbs?
- (c) If amla can be purchased at INR 30 per kg, should PNS buy amla?
- (d) What is the impact of increasing production of regular Chyawanprash by 1 kg on the optimal solution?
2. Powai Inc., a media marketing firm based in Mumbai, has contracted with a company to advertise its products. The company wants its TV and radio advertising to reach certain minimum number of customers within three age-groups: over 40, between 25 and 40, and under 25.

One minute of TV commercial time costs INR 90,000 and will reach (for every minute of advertisement) an average of 180,000 viewers in the over-40 group, 90,000 customers in the 25-to-40 group, and 120,000 in the under-25 group.

One minute of radio commercial time costs INR 25,000 and will reach (per every minute of advertisement) 40,000 listeners in the over-40 age-group, 80,000 in the 25-to-40 age-group, and 100,000 in the under-25 group.

The company wants to achieve a minimum exposure of 1000,000 in the over-40 group, 800,000 in the 25–40 age-group; the company did not have any requirement for the age group less than 25 years.

- (a) Formulate an appropriate linear programming model and solve it graphically to find the minimum cost that is required to meet the constraints. Identify the optimal number of minutes of advertisements in TV and radio.
- (b) By how much will the cost increase if the minimum exposure in the over-40 age group is increased by another 1000 people?
- (c) What is the maximum and minimum value of TV commercial cost per minute for which the optimal solution found in (a) remains optimal?
3. Easy Slim (ES) a manufacturer of health supplements is planning to introduce a miracle drink that will burn away body fat. ES guarantees that a person using health supplement will lose up to 10 pounds in just 3 weeks. The miracle drink can be made using 5 mystery ingredients A, B, C, D, and E. The plan calls for a person to consume at least 30 ounces of per week but not more than 50 ounces per week.

Each of the 5 ingredients A, B, C, D, and E contains different levels of three chemical components (X, Y, and Z). Health regulations mandate that dosage consumed per day should contain minimum prescribed levels of chemicals X and Y and should not exceed maximum prescribed level of chemical Z.

The composition of 5 ingredients in terms of the chemical components (units per ounce) is shown in Table 15.16 along with unit cost per ounce of the ingredients.

**TABLE 15.16**

Chemical	Ingredient					Minimum/Maximum Requirement
	A	B	C	D	E	
X	4	4	8	12	5	320 units
Y	5	5	8	10	12	300 units
Z	20	35	30	20	30	800 units
Cost per ounce	10	40	50	40	50	

Let  $A$ ,  $B$ ,  $C$ ,  $D$ , and  $E$  denote the number of ounces of ingredients A, B, C, D, and E to be used respectively to make the health drink. The corresponding linear programming formulation is given by

$$\text{Minimize cost } 10A + 40B + 50C + 40D + 50E \text{ (cost of daily dosage)}$$

subject to

$$A + B + C + D + E \geq 30 \text{ (daily dosage minimum)}$$

$$4A + 4B + 8C + 12D + 5E \geq 320 \text{ (chemical X requirement)}$$

$$5A + 5B + 8C + 10D + 12E \geq 300 \text{ (chemical Y requirement)}$$

$$20A + 35B + 30C + 20D + 30E \leq 800 \text{ (chemical Z max limit)}$$

$$A + B + C + D + E \leq 50 \text{ (daily dosage maximum)}$$

$$A, B, C, D, E \geq 0$$

The sensitivity output from Excel Solver is shown in Table 15.17.

**TABLE 15.17** Sensitivity report

Cell	Name	Final Value	Reduced Cost	Objective Coefficient	Allowable Increase	Allowable Decrease
\$A\$2	A	20	0	10	10	1E+30
\$B\$2	B	0		40	1E+30	45
\$C\$2	C	0		50	1E + 30	32

(Continued)

**TABLE 15.17** Sensitivity report—Continued

Cell	Name	Final Value	Reduced Cost	Objective Coefficient	Allowable Increase	Allowable Decrease
\$D\$2	D	20	0	40	8.888888889	20
\$E\$2	E	0	8	50	1E + 30	8
Cell	Name	Final Value	Shadow Price	Constraint RHS	Allowable Increase	Allowable Decrease
\$A\$4	Daily Dosage Minimum	40	0	30	10	1E + 30
\$A\$5	Chemical X Requirement	320		320	0	1E + 30
\$A\$6	Chemical Y Requirement	300		300	100	2.04836E-14
\$A\$7	Chemical Z Requirement	800	-1	800	1.63869E-13	200
\$A\$8	Daily Dosage Maximum	40	0	50	1E+30	10

- (a) Calculate the minimum cost of daily dosage.
- (b) Calculate the shadow price for constraint 2 (chemical X requirement) and constraint 3 (chemical Y requirement). Use complementary slackness theorem to answer.
- (c) What will be the impact of increasing ingredient B quantity in the health drink by one ounce on the cost of the drink? Use complementary slackness theorem to answer.
- (d) What will be the impact on current optimal solution if the cost of A increases to INR 15 (from 10) per ounce and cost of D increases to INR 44 (from 40) per ounce simultaneously? Will the optimal solution change? What will be the minimum cost of the health drink after the change in the costs of A and D?
- (e) A new ingredient, say ingredient F, is available at INR 20 per ounce. Ingredient F has 8 units of X, 12 units of Y, and 50 units of Z. Should ES use the new ingredient?
4. The Sidon company produces four types of alloys which we label 1, 2, 3, and 4. Each type of alloy (per gram) requires three different types of metals as shown Table 15.18.

**TABLE 15.18** Components of alloys

	Metal 1 (in grams)	Metal 2 (in grams)	Metal 3 (in grams)	Profit per gram
Alloy 1	0.2	0.4	0.4	186
Alloy 2	0.2	0.6	0.2	111
Alloy 3	0.3	0.3	0.4	281
Alloy 4	0.5	0.5	0	188

During the coming month, Sidon can acquire up to 2000 grams of metal 1, 3000 grams of metal 2, and 500 grams of metal 3. The unit costs are INR 5 per gram for metal 1, INR 5 per gram for metal 2, and INR 7 per gram for metal 3. The maximum demand for alloys 1, 2, 3, and 4 are 1000, 2000, 500, and 1000 grams respectively. The company wants to maximize its monthly profit by manufacturing the optimal quantity of alloys.

Let  $X_1, X_2, X_3$ , and  $X_4$  be the quantity of alloys (in grams) 1, 2, 3, and 4 to be manufactured. The corresponding LP formulation is given by

$$\text{Maximize } 186X_1 + 111X_2 + 281X_3 + 188X_4$$

subject to constraints

$$0.2X_1 + 0.2X_2 + 0.3X_3 + 0.5X_4 \leq 2000 \text{ (metal 1 constraint)}$$

$$\begin{aligned}
 0.4X_1 + 0.6X_2 + 0.3X_3 + 0.5X_4 &\leq 3000 \text{ (metal 2 constraint)} \\
 0.4X_1 + 0.2X_2 + 0.4X_3 &\leq 500 \text{ (metal 3 constraint)} \\
 X_1 &\leq 1000 \text{ (alloy 1 demand)} \\
 X_2 &\leq 2000 \text{ (alloy 2 demand)} \\
 X_3 &\leq 500 \text{ (alloy 3 demand)} \\
 X_4 &\leq 1000 \text{ (alloy 4 demand)} \\
 X_1, X_2, X_3, \text{ and } X_4 &\geq 0
 \end{aligned}$$

The Excel Solver output is provided Table 15.19.

**TABLE 15.19** Sensitivity report

Variable Cells						
Cell	Name	Final Value	Reduced Cost	Objective Coefficient	Allowable Increase	Allowable Decrease
\$A\$2	x1	0		186		1E+30
\$B\$2	x2		0	111	29.5	18
\$C\$2	x3		0	281	1E+30	59
\$D\$2	x4		0	188	1E+30	188
Constraints						
Cell	Name	Final Value	Shadow Price	Constraint RHS	Allowable Increase	Allowable Decrease
\$A\$5	Metal 1	950		2000	1E+30	1050
\$A\$6	Metal 2	1550		3000	1E+30	1450
\$A\$7	Metal 3	500		500	100	300
\$A\$8	Alloy 1	0		1000	1E+30	1000
\$A\$9	Alloy 2	1500		2000	1E+30	500
\$A\$10	Alloy 3	500		500	750	250
\$A\$11	Alloy 4	1000		1000	2100	1000

Use primal–dual relationship to answer the following questions:

- Write the optimal production plan (in terms of number of grams of alloys 1, 2, 3, and 4 manufactured). What is the objective function value?
- 200 additional grams of metal 3 can be imported and it would cost Sidon INR 20000. Should Sidon import this additional metal 3? State your answer clearly.
- What is the impact on the objective function value if the demand for alloy 3 is increased by 200 units?
- One of the customers of Sidon has placed order for alloy type 1. To maintain a long-term relationship Sidon would like to accept this order. How much should be the profit on alloy 1 so that Sidon can accept the order such that there will be no reduction from the current profit?
- One of the main competitors of Sidon goes bankrupt and using this opportunity, Sidon plans to increase the profits earned from different alloys to 279, 166.5, 421.5, and 282, respectively. State whether these changes to the profit will impact the current optimal production plan.
- Government imposes a restriction that Sidon can sell only 1000 grams of alloy 2. What will be the impact of this restriction on the optimal solution and the optimal profit?

- (g) A new alloy 5 can be manufactured using 0.5 grams of metal 1, 0.4 grams of metal 2, and 0.1 grams of metal 3 for one gram of alloy 5. The profit earned by alloy 5 is 220 per gram and the demand is 1500 grams per month. Should Sidon manufacture this new alloy?
5. Acharya Foods (AF) is a manufacturer of dosa (south Indian snacks) batter based out of Bangalore. The profit from 1 kg of dosa batter is INR 20; dosa batter is sold in a pack of 1 kg. During each production cycle, AF can make up to 500 kg of dosa batter per production cycle; however, there is a minimum requirement of 100 kg of batter to be manufactured every time the machine is used. The cost of operating the machine per cycle is INR 5000 and only two production cycles are possible on any given day. Any unsold batter will incur a loss of AF INR 30 (production and logistics cost). The demand for next seven days is shown in Table 15.20. The shelf life of batter is 3 days (including the day of manufacturing); any batter not sold within 3 days of its production is wasted.
- (a) Develop a multi-stage integer programming model to maximize the profit. Solve the problem to find the optimal production plan and optimal profit.
- (b) On day 4 the machine has to go through a preventive maintenance and it will not be available for producing dosa batter. Modify the formulation in (a) to incorporate this additional condition.

**TABLE 15.20** Demand for dosa batter for 7 days

Day	1	2	3	4	5	6	7
Demand	550	670	650	710	840	880	450

6. Solve the following binary integer programming problem using Branch and Bound Technique.

$$\text{Minimize } 4X_1 + 3X_2 + 5X_3 + 8X_4$$

subject to

$$6X_1 + 3X_2 + 5X_3 + 10X_4 \geq 12$$

$$X_i = 0 \text{ or } 1 \text{ for } i = 1, 2, 3, 4$$

7. Gayathri Iyer is an interior designer and provides interior design service to apartments in Bangalore. One of the frequently encountered problem is to minimize waste while cutting plywood into different sizes. The original size of the plywood is 6 feet long and 5 feet wide. For a specific customer the original plywood has to be cut into different units as per Table 15.21.

**TABLE 15.21** Required size of plywood along with demand for such size

S. No.	Size	Required Number (Demand)
1	4 × 4 feet	10
2	5 × 4 feet	16
3	3 × 6 feet	12
4	6 × 4 feet	8

- (a) Formulate an integer programming problem that will minimize the wastage in cutting the original plywood to make the required size as per the requirement in Table 15.21.
- (b) What is the optimal number of original size of plywood required to meet the demand in Table 15.21?
8. In the year 1002 AD, the King Raja Raja Chola laid foundation of the Brihadeeswarar Temple in Thanjavur. Raja Raja Rama was the architect and engineer of the temple. The temple is made of granite stones procured from various granite quarries of south India. To complete a section of the temple, Raja Raja Rama required 10,000 granite stones and the monthly requirements over 6 months starting from January 1002 is shown in Table 15.22

**TABLE 15.22** Monthly requirements of granite stones

Month	1	2	3	4	5	6
Blocks Required	1500	2500	2000	1200	2000	800

The granites are sourced from two main sources (say source 1 and source 2). Source 1 is about 50 km from Thanjavur and source 2 is about 60 km from Thanjavur. Elephants were used to move the blocks from both sources. Source 1 can supply 1200 blocks every month except for month 4. Source 2 can supply about 1800 blocks every month except month 5. The cost (2017 equivalent) of moving one block using elephant from source 1 is INR 8000 and from source 2 is INR 8500. Excess stones are stored in a nearby storage and its capacity is 200 blocks. The inventory cost of storing a granite block is INR 500.

- (a) Formulate a multi-stage integer programming problem to minimize the total cost of transporting granite stones to meet the monthly demand. Solve and find the optimal procurement plan.
  - (b) About 10% of granites become unusable due to damage while transporting them. Make necessary changes to the formulation in (a) to include this additional constraint. Solve and find the optimal solution.
9. Clara International Marketing (CIM) assists companies to advertise their products in television channels. The advertisement cost depends on the average television rating points (TRP). The programs, average TRP, cost of advertisement per minute are shown in Table 15.23.

**TABLE 15.23** TRP and cost of advertisement per minute

	Program				
	Cricket Matches	Other Sporting Event	Hindi Serials	Hindi Movies	English News Channels
Average TRP	4.2	3.5	2.8	2.5	0.2
Cost of advertisement per minute in (INR)	120,000	85,000	70,000	60,000	25,000

A customer of CIM has sent a proposal to promote their new product and has set the following goals:

1. The gross rating points (GRP) should be at least 100, where  $GRP = TRP \times \text{Number of Spots of 1 minute advertisement}$ .
2. The GRP through sports events should be at least 20.
3. The GRP for English news channels should not exceed 5.

The total budget for advertisement is INR 2 million. Develop a goal programming model that can be used for solving the problem; treat all three goals as equally important. Solve the problem using Excel Solver. The duration of the advertisement should be in integer multiple of minutes.

10. Metropolis multiplex has 4 movie screens. Each screen can show up to 4 movies in a day, that is total of 16 shows on any given day. The movie language and expected revenue from each show is shown in Table 15.24.

**TABLE 15.24** Movie Language and average revenue

S. No.	Movie Language	Revenue in INR Per Show
1	English	16000
2	Hindi	12400
3	Kannada	6200
4	Tamil	8900
5	Telugu	12100

The manager of the multiplex should plan the movies shows to meet the following goals:

- (a) There should be at least 4 shows of Kannada movies every day.
- (b) There should be at least one show of every language every day.
- (c) No language movie can have more than 8 shows in a day.
- (d) The total revenue per day should be at least 200000.

Formulate a goal programming problem to find the optimal screening of movies.

## Case Study

### Linen Management at Apollo Hospital<sup>1</sup>

It was a typical Bangalore evening, pleasant, a fresh breeze in the air, and intermittent short-lived monsoon showers, the ideal weather for sipping a hot cup of coffee. Admiring nature's exuberance through the *au bon pain* cafe's glass windows at the Indian Institute of Management, Bangalore (IIMB) campus, Dr Ananth N Rao, the head of the quality department at Apollo Hospital started explaining one of the operational problems that was perturbing him for the last few weeks. Sipping the hot coffee, Dr Rao said to Professor Dinesh Kumar:

In 2013, we spent around INR 5.1 million on the linen used in our hospital, about 67% was spent on washing alone. I believe that if managed optimally, we can reduce the money spent on linen significantly.

Apollo Hospital, Bangalore, has been a tertiary care flagship unit of the Apollo Hospitals Group accredited by the Joint Commission International (JCI). The hospital focused on centres of excellence such as cardiac sciences, neurosciences, orthopaedics, cancer, emergency medicine, and solid organ transplants, besides a complete range of more than 35 allied medical disciplines under the same roof.

The issue that was bothering Dr Rao in August 2014 was the process that was being followed at Apollo Hospital on Bannerghatta Road, Bangalore, for managing linen. The cost of managing linen (that is, the cost of buying new linen and of washing) was around INR 5.1 million per annum (\$1 = INR 60 in January 2014). They spent approximately INR 0.286 million per month on washing alone. The demand for linen was uncertain since many patients demanded more linen than was assigned to their bed for various reasons. To manage uncertain demand, Apollo maintained a safety stock for linen, since non-availability of linen could result in customer dissatisfaction. At the end of each day, the used linen was sent for different types of washing depending on the stains on the linen. A few heavily-soiled items of linen were discarded as they were not washable. Dr Rao discussed the linen management problem with his Chief Executive Officer and both of them were convinced that an analytics approach was required for effective management of linen at Apollo that would enable them to reduce the cost of linen management.

<sup>1</sup> The case was authored by Apoorva Sara Prakash, Muthu Solayappan and U Dinesh Kumar and is distributed through Harvard Business Publishing case portal. © 2015 Indian Institute of Management Bangalore. Reproduced with the permission of IIM Bangalore. This case is not intended to serve as an endorsement, source of primary data, or to show effective or inefficient handling of decision or business processes.

**Case Study** **Continued...**

## Linen in Hospitals

Linen is one of the essential requirements of a hospital, consisting of bed sheets, pillow covers, blankets, gowns, aprons, etc. Linen is used in all the departments of a hospital by the patients, doctors, nurses, and attendants. Soiled linen was sent to the laundry where it was washed with chemicals to get rid of any stains and contamination. Heavily-soiled linen was subjected to extra wash cycles. The cloth used in the linen met industry standards and was certified to handle a certain number of regular washes.

Linen in hospitals can shelter a large number of potentially harmful and disease-generating microorganisms.<sup>2</sup> This implies that an appropriate process should exist to avoid contamination of linen. Contamination of linen can lead to transmission of microorganisms to people and to the environment. All stages of linen management such as storage, handling, bagging, transporting, and washing have to be handled with care.<sup>3</sup>

Several standards and best practices have been generated by different health services across the world for storage, handling, and washing linen. Managing linen is important for hospitals since it can have a significant impact on the operating expenses of a hospital. Shortage of linen can result in patient dissatisfaction and lead to delay and cancellation of surgeries.<sup>4</sup>

## Linen Management Process

Apollo Hospital had a Central Linen Room where all the linen was stocked and distributed to various floors according to the demand from all the floors/units. Used linen was collected by the nurses and segregated into used and heavily-soiled linen. Heavily-soiled linen was collected separately in a yellow bag and thrown down the chute; the yellow bag could contain multiple linen items such as bed sheets, pillow covers, etc. Heavily-soiled linen went through a special cleaning process to ensure that it was cleaned and disinfected. The laundry representative counted the soiled linen and the yellow bags and issued a delivery challan (out). Linen was sent for washing every evening; it was washed in the laundry and returned to the Linen Room after two days (day + 2). The Linen Room maintained a book to record the number of linen items sent out and received. A certain percentage of bed sheets were also recycled into pillow covers. The linen use cycle is summarized in **Exhibit 1**.

The details of linen sent and received to/from the laundry were recorded in the delivery challan book. The linen issued to various floors was recorded in the Linen Room book. There was no established system to identify the number of washes before discarding an item based on the recommended number of washes for normal or heavily-soiled linen. The returned laundry was inspected for any stains or wear and tear. Approximately 0.28% of the used linen was discarded every day and about 40% of these discards were converted into pillowcases. Daily discards occurred mainly because of heavy stains. The Linen Room book was audited every morning. There were monthly audits to inspect the discarded linen and take account of the entire linen of the hospital. The flow of linen is summarized in **Exhibit 2**.

<sup>2</sup> "Guidelines for managing linen and laundry," Report, East Cheshire NHS Trust, December 2010. Document available at <http://www.eastcheshire.nhs.uk/About-The-Trust/policies/L/Linen%20and%20laundry%20management.pdf>

<sup>3</sup> Ibid.

<sup>4</sup> "Linen Management in Hospitals," *Clean India Journal*, February 24, 2012.

Case Study

Continued...

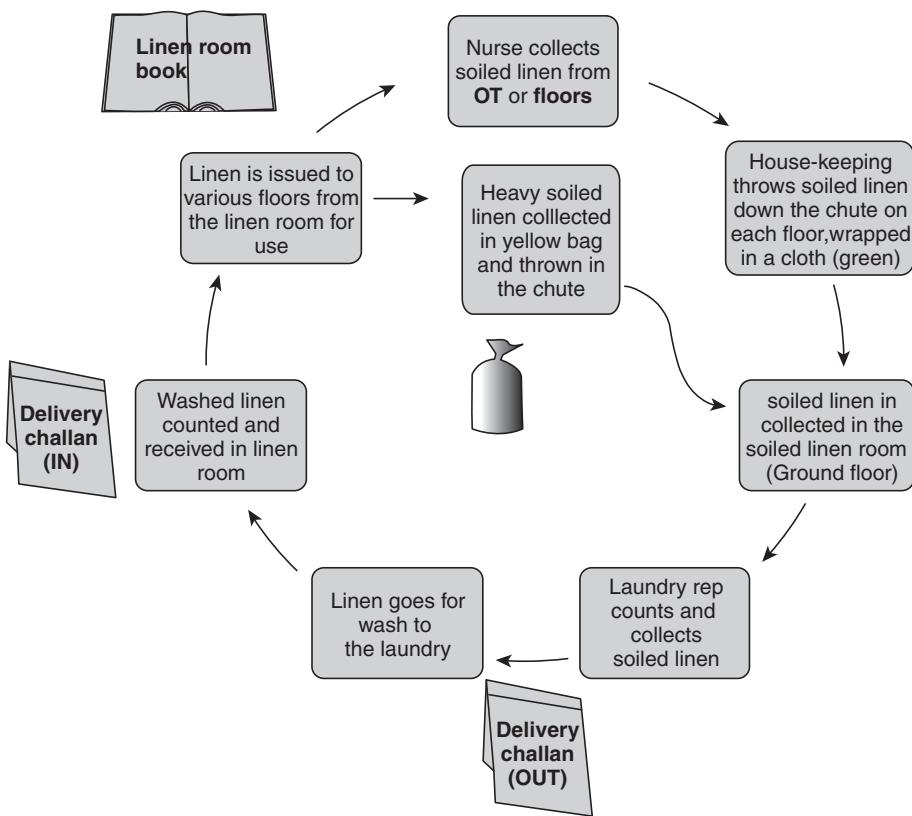


EXHIBIT 1 Linen use cycle. Source: Apollo Hospital.

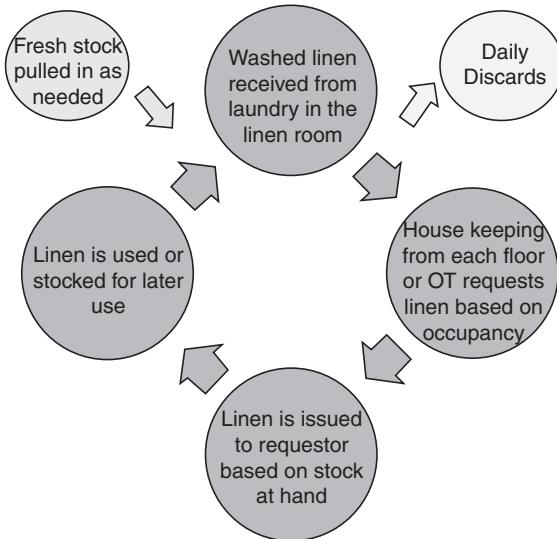


EXHIBIT 2 Linen flow cycle. Source: Apollo Hospital.

Continued...

### Occupancy Information

Linen usage depends on the occupancy rate of the hospital. The forecasted daily occupancy from October 2014 through December 2014 is shown in **Exhibit 3**.

**EXHIBIT 3** Forecasted daily occupancy data for October–December 2014

Day	October 2014	November 2014	December 2014
1	229	272	260
2	236	279	254
3	223	224	233
4	270	238	220
5	261	238	233
6	232	240	288
7	220	242	286
8	250	286	255
9	236	286	251
10	235	266	247
11	284	258	242
12	286	258	252
13	262	248	286
14	275	236	286
15	271	286	245
16	246	286	238
17	244	265	233
18	288	247	220
19	283	266	212
20	248	262	242
21	259	245	280
22	265	286	261
23	256	288	269
24	254	252	250
25	286	245	245
26	286	237	240
27	270	232	283
28	251	194	210
29	247	237	196
30	217	272	188
31	211	—	170

SOURCE: Apollo Hospital.

**Continued...**

Ananth added:

The hospital maintains a linen safety stock for 2 days (that is, 50% of the linen), 25% of the linen is under circulation (par count) and the remaining 25% is in the laundry.

Each bed had a defined par count (2 bed sheets and 1 pillow cover). The linen was changed every morning and also immediately after a patient was discharged. This increased the daily usage, which was accounted in the additional 2% used by the hospital.

### **Cost Information**

The cost of washing a unit of bed sheet was INR 5 and the cost of buying a new one was INR 181.5. Similarly, the cost of washing a pillow cover was INR 3.3 and the cost of buying a new one was INR 64.35.

### **The Challenge**

Apollo received approximately 250 patients every day and each occupancy required two bed sheets and a pillowcase (which is the par count). The capacity of the hospital for inpatients was 300 beds. The life of the linen depended on the number of washes: the linen could last up to approximately 70 wash cycles under regular wash; whereas the life reduced to 45 wash cycles for heavily-soiled linen. One of the challenges was to track the number of washes the linen had undergone. Technologies were available through radio frequency identification (RFID), using which linen usage and the number of cycles could be recorded accurately. However, RFID was not used by Apollo Hospital. According to Ananth Rao:

RFID cannot be implemented in hospitals in India as these RFIDs are costly and also may not survive the fluctuating washing patterns in India, thereby making the initial investment (even if made against better judgment) futile.

The cost of linen management through the year at Apollo was nearly INR 400 per month per bed for just the laundry expenses on bed linen. The total cost of managing linen per bed per month was INR 1,700 (which included all the linen such as bed sheets, pillow covers, gowns, aprons, etc.). The total cost of linen per month including procurement, replacement, maintaining two days' inventory, 0.28% discards, and laundry charges was approximately INR 4,25,000.

There was a need to relate the occupancy of inpatients to predict the number of bed sheets and pillow covers required on a daily basis. Apollo purchased linen once in a year. Dr Ananth Rao thought that they should increase the frequency of purchase. However, time between purchases should be at least 30 days to avoid too-frequent purchases. Apollo Hospital wanted to purchase an optimal quantity of linen at the right time to ensure that a sufficient amount of linen was available every day to

## Continued...

meet the demand as well as maintain a buffer inventory that would minimize the total cost. During its quarterly meeting held in August, the linen management committee of Apollo Hospital decided to buy a new stock of bed sheets and pillow covers for the quarter starting from October 2014 and destroy many old bed sheets and pillow covers. After destroying the old bed sheets and pillow covers, they expected that there would be 1,260 bed sheets and 680 pillow covers available on October 1, 2014. Also, Apollo Hospital was expecting 470 bed sheets and 240 pillow covers on October 1, 2014 and 460 bed sheets and 232 pillow covers on October 2, 2014 from the laundry.

Case Study

### CASE QUESTIONS

- Identify the decision variables that Dr Ananth N Rao should use for solving the linen management problem.
- Develop a mathematical programming model that can be used for minimizing the total cost of linen management at the Apollo Hospital for one quarter (October–December 2014).
- For Apollo, it costs INR 20,000 to employ a tailor who converts discarded bed sheets into pillow covers. Do you think that the decision by Apollo to convert bed sheets into pillow covers is cost-effective?
- Every day about 20% of the used linen goes through heavy wash and remaining through regular wash. Calculate the average life of the linen assuming that the life of linen is 70 washes under regular wash and 45 washes under heavy wash. Modify the formulation in question 2 to incorporate the fact that the leftover linen will be destroyed once it reaches the maximum number of allowable wash cycles.
- What are the limitations of the model and how should we handle these limitations?

### REFERENCES

- Armstrong A (2016), "Ocado Hits Milestone Delivering 250,000 Orders in a Week for the First Time", *The Telegraph*, 15 March 2016.
- Abraham P, Pradhan M, Lakshminarayanan Iyer G, and Kumar U D (2016), "Customer Analytics at Bigbasket – Product Recommendations", IIM Bangalore Case, IMB 573.
- Barnhart C, Belobaba P, and Odoni A R (2003). "Applications of Operations Research in Airline Industry", *Transportation Science*, 37(4), 368–391.
- Bradley S P, Hax A C, and Magnanti T L, "Applied Mathematical Programming", Addison Wesley Publishing Co, Reading.
- Charnes A, Cooper W W, and Ferguson R O (1955), "Optimal Estimation of Executive Compensation by Linear Programming", *Management Science*, 1(2), 138–151.
- Dantzig G (1963), *Linear Programming and Extensions*, Princeton University Press, Princeton.
- Dicken P (1977), "A Note on Location Theory and Large Business Enterprise", *Area*, 9(2), 138–143.
- Goldman A J and Tucker A W (1956), "Theory of Linear Programming", in *Linear Inequalities and Related Systems* (Eds. H W Kuhn and A W Tucker), Princeton University Press, Princeton.
- Lawler E L and Wood D E (1966), "Branch and Bound Methods: A Survey", *Operations Research*, 14(4), 699–719.
- Morell V (2012), "Flying Math. Bees Solve Travelling Salesman Problem", *Science Now*, 21 September 2012.
- Naylor T H (1966), "The Theory of Firm: A Comparison of Marginal Analysis and Linear Programming", *Southern Economic Journal*, 32(3), 263–274.
- Wendell R E (1985), "The Tolerance Approach to Sensitivity Analysis in Linear Programming", *Management Science*, 31(5), 564–578.



# Stochastic Models

16

“Problem Solving is often a matter of cooking up an appropriate Markov chain”.

— Olle Häggström

## LEARNING OBJECTIVES

- LO 16-1** Learn fundamental concepts in stochastic processes, different types of stochastic process models, and the concept of dependence in random variables.
- LO 16-2** Understand the need for collection of random variables in problem solving and decision making.
- LO 16-3** Learn Markov chain, properties of Markov chain, and its applications in solving analytics problems.
- LO 16-4** Learn Markov chain with absorbing states and its application across several business problems.
- LO 16-5** Understand customer lifetime value (CLV) and estimation of customer lifetime value using Markov chain.
- LO 16-6** Learn Markov decision process and its applications in solving sequential decision-making problems.
- LO 16-7** Learn Poisson process, compound Poisson process, and its applications.

## STOCHASTIC MODELS

Many systems are dynamic in nature, that is, use of single random variable is likely to be insufficient for modelling the problem and we may need a collection of random variables to model the problem effectively which is achieved through stochastic process models. Also, one of the major assumptions made in predictive analytics techniques discussed in Chapters 9–12 of this book is that all the random variables (outcome variables) are independent and identically distributed. However, this assumption may not be valid in many cases; random variables may have dependency relationship. In such cases, we may have to develop a model in which the random variables are dependent. Stochastic process models such as Markov process and semi-Markov process can be used for modelling problems in which the random variables are dependent. Markov decision process (MDP) is a reinforcement learning algorithm, an important class of machine learning algorithms. Reinforcement learning algorithms are used when both dependent and independent variables are uncertain.

**IMPORTANT**

*Stochastic models are powerful tools which can be used for solving problems which are dynamic in nature, that is, the values of the random variables change with time.*

## 16.1 | INTRODUCTION STOCHASTIC PROCESS

So far in this book we have used a single random variable (for example, in Chapters 9–12, outcome variable  $Y$  was the single random variable) to model various problems. Many real-life problems are dynamic in nature, that is, the values of key performance measures are likely to change with time. For example, market share of a brand will change with time and will depend on market share at previous time period. Thus, a single random variable may not be sufficient to model the problems in which the key performance measures change with time and may depend on several factors. Assume that an e-commerce retailer is interested in predicting monthly cash flow from its existing customers. The cash flow at time  $n$  can be represented using a collection of random variables  $\{X_n, \text{ where } n = 1, 2, \dots\}$ . That is, modelling the cash flow from the existing customers can be represented by a collection of random variables  $X_n$  observed at different time points. Stochastic process is defined as a collection of random variables  $\{X_n, n \geq 0\}$  indexed by time (however, index can be other than time). The value (cash flow) that the random variable  $X_n$  can take is called the **state of the stochastic process at time  $n$** . The set of all possible values the random variable can take is called the **state space**. If the system is observed over continuous time, then the stochastic process is written as  $\{X(t), t \geq 0\}$ .

The following are different stochastic process models classified based on certain properties:

1. Poisson Process
2. Markov Process and Markov Chain (MC)
3. Markov Decision Process
4. Partially Observable Markov Decision Process
5. Semi-Markov Process
6. Random Walk
7. Brownian Motion Process
8. Auto-Regressive and Moving Average Processes

In this chapter, we will be discussing Poisson process, compound Poisson process, Markov chains, and Markov decision process (MDP) with applications.

## 16.2 | POISSON PROCESS

In many cases, we would like to count the number of events that occur over a period of time. Following are few examples of counting process:

1. Retail stores would like to predict footfall (number of customers visiting the store) over a period of time.
2. Call centres would like to predict the number of calls they receive over a period of time.

3. Number of customer arrivals at banks, airports, restaurants, and any service centres.
4. Demand for spare parts of capital equipment caused due to failure of parts over a period of time.
5. Number of insurance claims received at an insurance company.

In the above examples, the primary objective is to count the number of events. Homogeneous Poisson Process (HPP) is a stochastic counting process  $N(t)$  with the following properties:

1.  $N(0) = 0$ , that is the number of events by time  $t = 0$  is zero.
2.  $N(t)$  has independent increments. That is if  $t_0 < t_1 < t_2 < \dots < t_n$ , then  $N(t_1) - N(t_0), N(t_2) - N(t_1), \dots, N(t_n) - N(t_{n-1})$  are independent.
3. The number of events by time  $t$ ,  $N(t)$ , follows a Poisson distribution, that is

$$P[N(t)=n] = \frac{e^{-\lambda t} \times (\lambda t)^n}{n!} \quad (16.1)$$

Cumulative distribution of number of events by time  $t$  in a Poisson process is given by

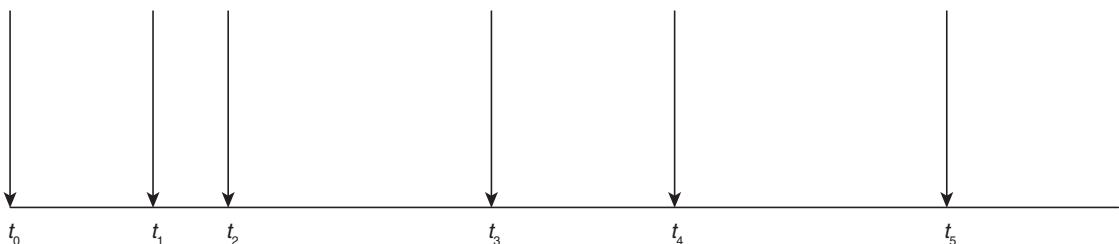
$$P[N(t) \leq n] = \sum_{i=0}^n P[N(t)=i] = \sum_{i=0}^n \frac{e^{-\lambda t} \times (\lambda t)^i}{i!} \quad (16.2)$$

The mean,  $E[N(t)]$ , and variance,  $\text{Var}[N(t)]$ , of a Poisson process  $N(t)$  are given by

$$E[N(t)] = \lambda t \quad (16.3)$$

$$\text{Var}[N(t)] = \lambda t \quad (16.4)$$

Figure 16.1 shows a Poisson process of events in which  $(t_1 - t_0), (t_2 - t_1), (t_3 - t_2)$  are time between events. In the case of Poisson process, the time between events follows an exponential distribution with parameter  $\lambda$ , that is the time between events have a density function  $f(t) = \lambda e^{-\lambda t}$  and cumulative distribution function  $F(t) = 1 - e^{-\lambda t}$ .



**FIGURE 16.1** Poisson process.

### EXAMPLE 16.1

Johny Sparewala (JS) is a supplier of aircraft flight control system spares based out of Mumbai, India. The demand for hydraulic pumps used in the flight control system follows a Poisson process. Sample data (50 cases) on time between demands (measured in number of days) for hydraulic pumps are shown in Table 16.1.

**TABLE 16.1** Time between demands (in days) for hydraulic pumps

104	90	45	32	12	6	30	23	58	118
80	12	216	71	29	188	15	88	88	94
63	125	108	42	77	65	18	25	30	16
92	114	151	10	26	182	175	189	14	11
83	418	21	19	73	31	175	14	226	8

- (a) Calculate the expected number of demand for hydraulic pump spares for next two years.
- (b) Johny Sparewala would like to ensure that the demand for spares over next two years is met in at least 90% of the cases from the spares stocked (called fill rate) since lead time to manufacture a part is more than 2 years. Calculate the inventory of spares that would give at least 90% fill rate.

**Solution:**

- (a) To calculate the expected number of demand for spares for two years, we have to estimate the parameter  $\lambda$  of the Poisson distribution. The maximum likelihood estimate of  $\lambda$  is given by

$$\hat{\lambda} = \frac{1}{\frac{1}{n} \sum_{i=1}^n X_i} = 0.0125$$

where  $X_i$  is the time between failure of  $i^{\text{th}}$  case and  $\frac{1}{n} \sum_{i=1}^n X_i$  is the mean time between failure.

The expected number of demand for spares,  $E[N(t)]$ , for 2 years ( $2 \times 365$  days) is given by

$$E[N(t)] = E[N(2 \times 365)] = \hat{\lambda} \times t = 0.0125 \times 2 \times 365 = 9.125$$

- (b) To ensure that the demand for spares is met 90% of the time, we have to calculate smallest  $k$  such that

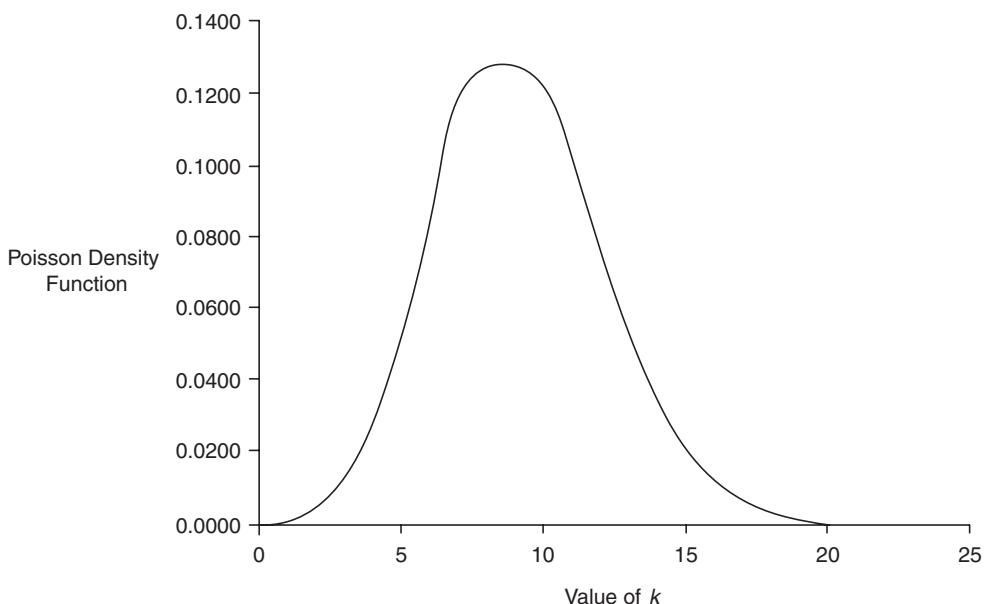
$$\sum_{i=0}^k \frac{e^{-\hat{\lambda}t} \times (\hat{\lambda}t)^i}{i!} \geq 0.90$$

Table 16.2 shows density and cumulative distribution function values of Poisson process for different values of  $k$ .

**TABLE 16.2** Poisson density and distribution function for different values of  $k$ 

$k$	Poisson Density	Cumulative	$k$	Poisson Density	Cumulative
0	0.0001	0.0001	11	0.0996	0.7907
1	0.0010	0.0011	12	0.0758	0.8665
2	0.0045	0.0056	13	0.0532	0.9197
3	0.0138	0.0194	14	0.0347	0.9543
4	0.0315	0.0509	15	0.0211	0.9754
5	0.0574	0.1083	16	0.0120	0.9875
6	0.0873	0.1956	17	0.0065	0.9939
7	0.1138	0.3095	18	0.0033	0.9972
8	0.1298	0.4393	19	0.0016	0.9988
9	0.1316	0.5709	20	0.0007	0.9995
10	0.1201	0.6911	21	0.0003	0.9998

Smallest value of  $k$  for which the cumulative probability is greater than 0.90 is 13. That is, JS should stock 13 spares to ensure that they meet demand for spares in 90% of the cases over a two-year period. The probability density function of Poisson distribution with mean 9.125 is shown in Figure 16.2.

**FIGURE 16.2** Poisson process density function.

### 16.3 | COMPOUND POISSON PROCESS

Consider a Poisson process which is the arrival of customers to withdraw cash from an automatic teller machine (ATM). Every arriving customer is likely to withdraw cash the amount of cash withdrawn is likely to be different for different customers. Assume that cash withdrawn by customers follows independent and identically distributed random variables  $Y_i$ , where  $Y_i$  is the cash withdrawn by  $i^{\text{th}}$  customer. The bank would be interested in predicting the total amount of cash withdrawn over a period of time for effective cash replenishment.

Compound Poisson process is a stochastic process  $X(t)$  where the arrival of events follows a Poisson process and each arrival is associated with another independent and identically distributed (IID) random variable  $Y_i$ . Compound Poisson process  $X(t)$  is a continuous-time stochastic process defined as

$$X(t) = \sum_{k=1}^{N(t)} Y_k \quad (16.5)$$

where  $N(t)$  is a Poisson process with mean  $\lambda t$  and  $Y_i$  are IID random variables with mean  $E(Y_i)$  and variance  $\text{Var}(Y_i)$ .

The mean and variance of the compound Poisson process  $X(t)$  are given by (Ross, 2010)

$$E[X(t)] = \mu_{X(t)} = \lambda t \times E(Y_i) \quad (16.6)$$

$$\text{Var}[X(t)] = \sigma_{X(t)}^2 = \lambda t \times E(Y_i^2) = \lambda t \times (\text{Var}(Y_i) + [E(Y_i)]^2) \quad (16.7)$$

For large  $t$ , we can show that the compound Poisson process follows an approximate normal distribution with mean  $\mu_{X(t)}$  and standard deviation  $\sigma_{X(t)}$ .

#### EXAMPLE 16.2

Customers arrive at an average rate of 12 per hour to withdraw money from an ATM and the arrivals follow a Poisson process. The money withdrawn are independent and identically distributed with mean and variance INR 4200 and 2,50,000, respectively. If the ATM has INR 6,00,000 cash, what is the probability that it will run out of cash in 10 hours?

#### Solution:

The mean and standard deviation of the compound Poisson process  $X(t)$  can be calculated as described below:

Mean of compound Poisson process is

$$\mu_{X(t)} = \lambda t \times E(Y_i) = 12 \times 10 \times 4200 = 5,04,000$$

Variance of compound Poisson process is

$$\sigma_{X(t)}^2 = \lambda t \times (\text{Var}(Y_i) + [E(Y_i)]^2) = 12 \times 10 \times (250000 + 4200^2) = 21468 \times 10^5$$

Standard deviation of compound Poisson process is

$$\sigma_{X(t)} = \sqrt{\sigma_{X(t)}^2} = \sqrt{21468 \times 10^5} = 46333.57$$

Probability that the cash withdrawal will exceed INR 6,00,000 is given by

$$P(X(t) \geq 6,00,000) = P\left(Z \geq \frac{6,00,000 - 504000}{46333.57}\right) = P(Z \geq 2.0719) = 0.0191$$

That is, there is approximately 2% chance that the ATM will run out of cash in 10 hours.

## 16.4 | MARKOV CHAINS

Let  $\{X_n, n = 0, 1, 2, \dots\}$  be a sequence of random variables observed at time  $n (= 0, 1, 2, \dots)$ . For example,  $X_n$  can be a market share of a company at time  $n$  or number of customers waiting at a supermarket checkout counter at time  $n$  or value of a share at time  $n$ . The value of  $X_n$  at time  $n$  is called the **state** and the set of all possible values is called the **state space** ( $S$ ). In the case of market share, the state space is a real number between 0 and 100 (between 0 and 100 percentage) and in the case of customers waiting at the checkout counter, the state space will be an integer, that is,  $S = \{0, 1, 2, \dots\}$ . We would like to calculate the conditional probability for the value of the random variable  $X_{n+1}$  given the history,  $P[X_{n+1} = j | X_0 = i_0, X_1 = i_1, \dots, X_n = i]$ , that is, we would like to predict the future state of the process  $(X_{n+1})$  given the history  $(X_n, X_{n-1}, \dots, X_0)$ . The process  $\{X_n, n = 0, 1, 2, \dots\}$  is a **first-order Markov process** if

$$P[X_{n+1} = j | X_0 = i_0, X_1 = i_1, \dots, X_n = i] = P[X_{n+1} = j | X_n = i] \quad (16.8)$$

The condition (16.8) is called Markov property named after the Russian mathematician A A Markov. If the state space  $S$  is discrete then the stochastic process  $\{X_n, n = 0, 1, 2, \dots\}$  that satisfies the condition (16.8) is called a Markov chain. If the state space is discrete and the process is observed over a continuous time and satisfies Eq. (16.8), then the process  $\{X(t), t \geq 0\}$  is called a continuous-time Markov Chain (CTMC). If the state space is continuous then the stochastic process that satisfies the Markov property is called a Markov process.

Markov chain is one of the most powerful tools in analytics and is successfully deployed by many companies. For example, the famous PageRank algorithm (Brin and Page, 1998; Hayes, 2013) used by Google for ranking web pages is based on Markov chain model. Google's Page ranking algorithm was one of the reasons why it was successfully able to compete with other search engines such as Alta Vista and Infoseek (which do not exist today) during the initial days of Google. An earlier successful solution based on Markov chain (actually based on Markov decision process) was proposed by Ronald Howard (Howard, 2002) to arrive at a catalog mailing policy for the American retail giant Sears in 1980s. As commented by Olle Häggström (2007), "*Problem Solving is often a matter of cooking up an appropriate Markov chain*". The following are few cases of endless applications of Markov chain:

1. Predicting credit ratings of companies are often modelled using Markov chain, where the states are the different ratings (such as AAA, AA, etc.).

2. Movement of stock price in which the future value of the stock price is often modelled using Markov Chain.
3. Condition of assets such as aircraft and capital equipment over a period of time can be modelled as a Markov chain.
4. Predicting brand switching and market share of different brands over time.
5. Prediction of Non-Performing Assets (bad loan) in banking and finance sector.
6. Any sequential decision-making scenario such as whether to buy or sell shares during a planning horizon.
7. Spell check and sentiment analysis is often modelled using hidden Markov models (HMM).

In the next few sections, we will be discussing the fundamental concepts in Markov chain along with its applications.

#### 16.4.1 | One-Step Transition Probabilities of Markov Chain

Let  $\{X_n, n = 0, 1, 2, \dots\}$  be a Markov chain with  $m$  states. Then the conditional probability

$$P[X_{n+1} = j | X_0 = i_0, X_1 = i_1, \dots, X_n = i] = P[X_{n+1} = j | X_n = i] = P_{ij} \quad (16.9)$$

is called the one-step transition probability.  $P_{ij}$  gives conditional probability of moving from state  $i$  to stage  $j$  in one period. One-step transition probabilities between all states in the state space are expressed in the form of one-step transition probability matrix as shown in Eq. (16.10).

$$\mathbf{P} = \mathbf{P}_{ij} = \begin{pmatrix} & 1 & 2 & \dots & m \\ 1 & P_{11} & P_{12} & \dots & P_{1m} \\ 2 & P_{21} & P_{22} & \dots & P_{2m} \\ \dots & \dots & \dots & \dots & \dots \\ m & P_{m1} & P_{m2} & \dots & P_{mm} \end{pmatrix} \quad (16.10)$$

An  $s$ -step transition probability in a Markov chain is given by

$$P_{ij}^{(s)} = P(X_{n+s} = j | X_n = i) \quad (16.11)$$

Equation (16.11) provides the probability of reaching state  $j$  from state  $i$  in exactly  $s$  steps ( $s$  periods).

The  $s$ -step transition probability  $P_{ij}^{(s)}$  can be written as

$$P_{ij}^{(s)} = \sum_{r=1}^m P_{ir}^k \times P_{rj}^{(s-k)}, \quad 0 < k < s \quad (16.12)$$

The relationship (16.12) is due to Chapman and Kolmogorov and is known as Chapman–Kolmogorov equation (Cinlar, 1975). Estimating one-step probability transition matrix is an important step in Markov chain model building which will be discussed in the next section.

### 16.4.2 | Estimation of One-Step Transition Probabilities of Markov Chain

Transition probabilities of a Markov chain are estimated using maximum likelihood estimate (MLE) from the transition data (Anderson and Goodman, 1957). Assume that the Markov chain has  $m$  states. Then the MLE estimate of the transition probability  $P_{ij}$  (probability of moving from state  $i$  to state  $j$  in one step) is given by

$$\hat{P}_{ij} = \frac{N_{ij}}{\sum_{k=1}^m N_{ik}} \quad (16.13)$$

where  $N_{ij}$  is number of cases in which  $X_n = i$  (state at time  $n$  is  $i$ ) and  $X_{n+1} = j$  (state at time  $n + 1$  is  $j$ ). Alternatively,  $N_{ij}$  can be interpreted as number of observations at state  $i$  at time  $n$  moving to state  $j$  at time  $n + 1$ . To estimate  $P_{ij}$ , transition data between states is collected over several periods and the one-step transition matrix is calculated using the data obtained from several periods using Eq. (16.13).

### 16.4.3 | Hypothesis Tests for Markov Chain: Anderson Goodman Test

Before we use the estimated values of the one-step transition probabilities [Eq. (16.13)], we need to check whether the sequence of random variables,  $X_n$ , form a Markov chain. This is carried out using Anderson–Goodman test (1957) which is a chi-square test of independence. The null and alternative hypotheses to check whether the sequence of random variables follows a Markov chain is stated below:

- $H_0$ : The sequences of transitions  $(X_1, X_2, \dots, X_n)$  are independent (zero-order Markov chain)
- $H_A$ : The sequences of transitions  $(X_1, X_2, \dots, X_n)$  are dependent (first-order Markov chain)

The corresponding test statistic is

$$\chi^2 = \sum_{i=1}^m \sum_{j=1}^m \left( \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right) \quad (16.14)$$

where

$O_{ij}$  = Observed number of transitions from state  $i$  to state  $j$  in one period.

$E_{ij}$  = Expected number of transitions from state  $i$  to state  $j$  assuming independence.

An excellent example of estimation of transition probabilities and checking for whether the data follows a Markov chain is discussed by Styan and Smith (1964). Example 16.3 is motivated by the example presented by them.

#### EXAMPLE 16.3

Coonoor and Co (CoCo) is a chain of retail stores in Nilgiris district of Tamil Nadu. CoCo sells two brands of rusks (hard biscuit): (a) Coonoor Lite (CL) and (b) Ooty Strong (OS). CoCo maintains an ERP system that captures the point of sale (POS)

data of items purchased by its customers. Since most of CoCo customers have loyalty cards, it can track the purchase behaviour of its customers. CoCo decided to track the purchase behaviour of 250 customers who had loyalty card. Since rusk is a daily snacks item, CoCo believed that most of its customers would buy rusk at least once in a week. Data (Table 16.3) was collected over 6 weeks using the states defined as follows:

**STATE 1**

Customer purchased the brand Coonoor Lite

**STATE 2**

Customer purchased the brand Ooty Strong

**STATE 3**

Customer purchased both Coonoor Lite and Ooty Strong

**STATE 4**

Customer did not buy any of the brands

**TABLE 16.3** Customer purchase behaviour

Week	State 1	State 2	State 3	State 4
1	145	47	40	18
2	133	46	36	35
3	125	57	34	34
4	116	58	38	38
5	116	58	35	41
6	109	62	37	42

The data in Table 16.3 has to be interpreted as follows: During week 1, 145 customers were in state 1, 47 in state 2, 40 in state 3, and 18 in state 4. The frequency transition matrix  $F_{t,t+1}$  is shown in Table 16.4. The frequency transition between week 1 and 2 has to be interpreted as follows: During week 1, 145 customers were in state 1. Out of these 145 customers, 116 customers purchased CL (remained in state 1), 15 purchased OS (moved to state 2), 7 purchased both (moved to state 3), and 7 did not purchase any (moved to state 4). That is, from the frequency transition matrix  $F_{12}$  in Table 16.4, we know that 116 customers stayed in state 1, 15 moved to state 2, 7 moved to state 3, and 7 moved to state 4. The corresponding estimate of the transition probabilities based on frequency of transition between week 1 and week 2 are given by

$$\hat{P}_{11} = \frac{116}{145} = 0.800, \hat{P}_{12} = \frac{15}{145} = 0.103, \hat{P}_{13} = \frac{7}{145} = 0.048, \text{ and } \hat{P}_{14} = \frac{7}{145} = 0.048$$

**TABLE 16.4** Frequency transition matrix

Frequency of transition between weeks 1 and 2					Frequency of transition between weeks 2 and 3				
$F_{12}$	1	2	3	4	$F_{23}$	1	2	3	4
1	116	15	7	7	1	108	13	7	5
2	8	25	3	11	2	6	35	4	1
3	8	4	24	4	3	8	4	21	3
4	1	2	2	13	4	3	5	2	25
Frequency of transition between weeks 3 and 4					Frequency of transition between weeks 4 and 5				
$F_{34}$	1	2	3	4	$F_{45}$	1	2	3	4
1	102	11	6	6	1	100	6	4	6
2	5	43	6	3	2	5	44	6	3
3	7	3	21	3	3	8	4	22	4
4	2	1	5	26	4	3	4	3	28
Frequency of transition between weeks 5 and 6									
$F_{56}$	1	2	3	4					
1	94	11	6	5					
2	6	44	5	3					
3	7	3	22	3					
4	2	4	4	31					

The one-step transition probability matrix based on the frequency table between week 1 and week 2 is given by

$$\mathbf{P} = \begin{pmatrix} & 1 & 2 & 3 & 4 \\ 1 & 0.800 & 0.103 & 0.048 & 0.048 \\ 2 & 0.170 & 0.532 & 0.064 & 0.234 \\ 3 & 0.200 & 0.100 & 0.600 & 0.100 \\ 4 & 0.056 & 0.111 & 0.111 & 0.722 \end{pmatrix} \quad (16.15)$$

Note that we will end-up with 5 different transition matrices when we use the transition frequency matrices provided in Table 16.4. Alternatively, we can consolidate the entire data in Tables 16.3 and 16.4 to estimate transition probability values based on aggregate data. In the first 5 weeks, the total number of people who purchased brand 1 (CL) is 635, out of which 520 again purchased brand 1 (CL). Thus, the estimated probability based on the aggregate data is

$$\hat{P}_{11} = \frac{520}{635} = 0.8189$$

We can calculate estimated probability values for other states. The one-step transition matrix based on aggregated data is shown in Eq. (16.16).

$$\mathbf{P} = \begin{pmatrix} & 1 & 2 & 3 & 4 \\ 1 & 0.8189 & 0.0882 & 0.0472 & 0.0457 \\ 2 & 0.1128 & 0.7180 & 0.0902 & 0.0789 \\ 3 & 0.2077 & 0.0984 & 0.6011 & 0.0929 \\ 4 & 0.0663 & 0.0964 & 0.0964 & 0.7410 \end{pmatrix} \quad (16.6)$$

Before we use the transition matrix  $\mathbf{P}$  to gain insights, we have to check whether the transitions follow a first-order Markov chain. That is, we have to conduct the chi-square of independence to check whether the transitions follow a first-order Markov chain. Table 16.5 shows the chi-square test results for the 5 frequency transition matrices in Table 16.4.

**TABLE 16.5** Chi-square results for testing whether the transitions follow a first-order Markov chain

Transition Frequency Matrix	Chi-square statistic value	p-value ( $df=9$ )
$F_{12}$	210.1368	$2.47 \times 10^{-40}$
$F_{23}$	289.3521	$4.72 \times 10^{-57}$
$F_{34}$	310.6549	$1.43 \times 10^{-61}$
$F_{45}$	318.9998	$2.41 \times 10^{-63}$
$F_{56}$	315.3137	$1.46 \times 10^{-62}$

Based on the  $p$ -values in Table 16.5, we reject the null hypothesis (random variables are independent), that is, we can accept that the transitions are first-order Markov chain. Chi-square statistic value for the consolidated frequency data is 1434.82 and the corresponding  $p$ -value is  $2.3149 \times 10^{-303}$ . Thus, we have strong evidence to suggest that the transitions follow a Markov chain. Note that the values of transition probability obtained using individual frequency transition matrices and consolidated data over 6 weeks are different. We have to check whether the transition matrices are time homogeneous before using the transition matrix for prediction and modelling. The time homogeneity of transition matrix is tested using likelihood ratio test discussed in the next section.

### 16.4.4 | Testing Time Homogeneity of Transition Matrices: Likelihood Ratio Test

In Example 16.3, we will have 5 different transition probability matrices (5 different estimates of  $P_{ij}$ ) when we use the transition frequency matrices in Table 16.4. We can estimate the values of  $P_{ij}$  using aggregate data in Tables 16.3 and 16.4 as shown in Eq. (16.16). Notice that the estimate for  $P_{11}$  based on transitions between weeks 1 and 2 is 0.800, whereas the estimate based on the aggregate data (all 6 weeks) is 0.818. We have to check for time homogeneity of the transition probability values before using the matrix for further analysis of the system. Anderson and Goodman (1957) suggested a likelihood ratio test for checking whether the transition probability matrices are time homogeneous. The null and alternative hypotheses of the likelihood ratio tests are

$$\begin{aligned} H_0: P_{ij}(t) &= \hat{P}_{ij}, t = 1, 2, 3, 4, \text{ and } 5 \\ H_A: P_{ij}(t) &\neq \hat{P}_{ij}, t = 1, 2, 3, 4, \text{ and } 5 \end{aligned}$$

where  $P_{ij}(t)$  is the estimated value of transition probability between state  $i$  and  $j$  based on the data for periods  $t$  and  $t + 1$ .  $\hat{P}_{ij}$  is the hypothesis value, which is value calculated based on the aggregate data,  $n_i(t)$  in Eq. (16.17) is the frequency of transition between states  $i$  and  $j$  at time  $t$ . The test statistic is a likelihood ratio test statistic and is given by (Anderson and Goodman, 1957):

$$\lambda = \prod_t \prod_{i,j} \left[ \frac{\hat{P}_{ij}}{\hat{P}_{ij}(t)} \right]^{n_{ij}(t)} \quad (16.17)$$

The test statistic in Eq. (16.17) is equivalent to (Anderson and Goodman, 1957)

$$\chi^2 = \sum_t \sum_i \sum_j \frac{n_i(t) \left[ \hat{P}_{ij}(t) - \hat{P}_{ij} \right]^2}{\hat{P}_{ij}} \quad (16.18)$$

where  $n_i(t)$  is the number of customers in state  $i$  at time  $t$ . The test statistic in Eq. (16.18) follows a  $\chi^2$  distribution with  $(t - 1) \times m \times (m - 1)$  degrees of freedom (Anderson and Goodman, 1957). For Example 16.3, the test statistic for time homogeneity of transition matrices is given by

$$\chi^2 = \sum_{t=1}^5 \sum_{i=1}^4 \sum_{j=1}^4 \left( \frac{n_i(t) [\hat{P}_{ij}(t) - \hat{P}_{ij}]^2}{\hat{P}_{ij}} \right) \quad (16.19)$$

The corresponding value of chi-square statistic is 32.1540, the critical chi-square value at  $\alpha = 0.05$  is 65.1707, and the corresponding  $p$ -value ( $df = 48$ ) is 0.9616. Since the  $p$ -value is high, we retain the null hypothesis, that is, the transition probability matrices are time homogeneous.

### 16.4.5 | Using Markov Chains in Predictive Analytics

One of the primary applications of Markov chain is predicting the values of  $X_n$  in the future. For example, assume that the initial distribution of customers in 4 states of Example 16.3 is  $\mathbf{P}_I = (450, 225, 175, 150)$ . Assume that the one-step transition matrix  $\mathbf{P}$  is as shown in Eq. (16.16).

Using Chapman–Kolmogorov relationship [Eq. (16.12)], we can show that the distribution of customers after  $n$  periods is given by  $\mathbf{P}_I \times \mathbf{P}^n$ , where  $\mathbf{P}_I$  is the initial distribution of customers across various states and  $\mathbf{P}$  is the one-step transition matrix. For example, the distribution of customers after 4 weeks is  $\mathbf{P}_I \times \mathbf{P}^4$ . That is

$$(450 \ 225 \ 175 \ 150) \times \begin{pmatrix} 0.8189 & 0.0882 & 0.0472 & 0.0457 \\ 0.1128 & 0.7180 & 0.0902 & 0.0789 \\ 0.2077 & 0.0849 & 0.6011 & 0.0929 \\ 0.0663 & 0.0964 & 0.0964 & 0.7410 \end{pmatrix}^4 = (417.84 \ 243.06 \ 150.69 \ 188.41)$$

So after 4 periods, the distribution of customers will be

State 1: 417.84; State 2: 243.06; State 3: 150.69; and State 4: 188.41

#### 16.4.6 | Stationary Distribution in a Markov Chain

For large value of  $n$ , value of  $\mathbf{P}_I \times \mathbf{P}^n$  may converge to a value which is known as the stationary distribution (also known as steady-state distribution) of the Markov chain and is denoted as  $\pi = \{\pi_1, \pi_2, \dots, \pi_m\}$ . Consider brand switching between two brands ( $B_1$  and  $B_2$ ) and let the initial market share be as shown in the following vector:

$$\mathbf{P}_I = (0.2 \ 0.8)$$

Let the one-step transition probability matrix between two brands be as shown in Table 16.6.

TABLE 16.6 Transition probability

	Brand 1	Brand 2
Brand 1	0.80	0.20
Brand 2	0.25	0.75

Transition probability in Table 16.6 can be represented using the state transition diagram shown in Figure 16.2.

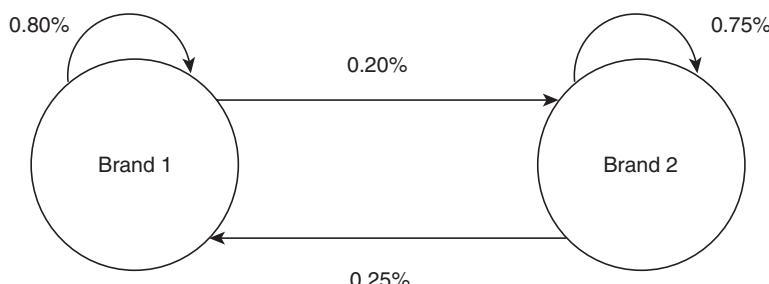


FIGURE 16.3 State transition diagram between brands.

**TABLE 16.7** Shows the values of  $\mathbf{P}^n$  and the market share of brands after  $n$  periods ( $\mathbf{P}_1 \mathbf{P}^n$ )

	<b>Brand 1</b>	<b>Brand 2</b>	<b>Market Share <math>n</math> Periods</b>		
$\mathbf{P}^1$	0.2	0.8			
	Brand 1	Brand 2		Brand 1	Brand 2
P	Brand 1	0.8	0.2	1 ( $\mathbf{P}_1 \mathbf{P}^1$ )	0.36
	Brand 2	0.25	0.75		0.64
$\mathbf{P}^2$	Brand 1	0.69	0.31	2 ( $\mathbf{P}_1 \mathbf{P}^2$ )	0.448
	Brand 2	0.3875	0.6125		0.552
$\mathbf{P}^4$	Brand 1	0.596225	0.403775	4 ( $\mathbf{P}_1 \mathbf{P}^4$ )	0.52302
	Brand 2	0.504719	0.495281		0.47698
$\mathbf{P}^8$	Brand 1	0.559277	0.440723	8 ( $\mathbf{P}_1 \mathbf{P}^8$ )	0.552578
	Brand 2	0.550904	0.449096		0.447422
$\mathbf{P}^{16}$	Brand 1	0.555587	0.444413	16 ( $\mathbf{P}_1 \mathbf{P}^{16}$ )	0.555531
	Brand 2	0.555517	0.444483		0.444469
$\mathbf{P}^{32}$	Brand 1	0.555556	0.444444	32 ( $\mathbf{P}_1 \mathbf{P}^{32}$ )	0.555556
	Brand 2	0.555556	0.444444		0.444444
$\mathbf{P}^{64}$	Brand 1	0.555556	0.444444	64 ( $\mathbf{P}_1 \mathbf{P}^{64}$ )	0.555556
	Brand 2	0.555556	0.444444		0.444444

In Table 16.7, both rows of the matrix  $\mathbf{P}^n$  converge to 0.555556 and 0.444444 as the value of  $n$  increases. The market share of brands 1 and 2 converges to 0.555556 and 0.444444, respectively. The values (0.555556, 0.444444) are the stationary probability distribution of the Markov chain or equilibrium probabilities. The values can be interpreted as long-run market shares of the brands.

Let  $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_m)$  be the stationary distribution. Then it satisfies the following system of equations:

$$\pi_j = \sum_{k=1}^m \pi_k P_{kj} \quad (16.20)$$

$$\sum_{k=1}^m \pi_k = 1 \quad (16.21)$$

The system of equations in Eq. (16.20) can be written as

$$\boldsymbol{\pi} = \boldsymbol{\pi} \mathbf{P} \quad (16.22)$$

The stationary distribution equation for the matrix in Table 16.6 is given by

$$(\pi_1 \quad \pi_2) = (\pi_1 \quad \pi_2) \begin{pmatrix} 0.80 & 0.20 \\ 0.25 & 0.75 \end{pmatrix} \quad (16.23)$$

That is

$$\pi_1 = 0.80\pi_1 + 0.25\pi_2 \quad (16.24)$$

$$\pi_2 = 0.20\pi_1 + 0.75\pi_2 \quad (16.25)$$

Since  $\pi_1$  and  $\pi_2$  are probabilities, we have

$$\pi_1 + \pi_2 = 1 \quad (16.26)$$

To find the solution to the system of equations in Eqs. (16.24)–(16.26), we have to always use  $\pi_1 + \pi_2 = 1$  and one of the equations between Eq. (16.24) and Eq. (16.25). Equation (16.24) can be written as

$$\begin{aligned} 0.20\pi_1 - 0.25\pi_2 &= 0 \\ \pi_1 + \pi_2 &= 1 \end{aligned}$$

Solving the above system of equations, we get  $\pi_1 = 0.555556$  and  $\pi_2 = 0.444444$ . That is, in the long run, the markets shares of brand 1 and brand 2 will converge to 0.555556 and 0.444444, respectively. The stationary distribution will be independent of the initial probability distribution  $P_r$ .

**IMPORTANT**

*The stationary distribution, if exists, will be independent of the initial distribution  $P_r$*

#### 16.4.7 | Regular Matrix

A matrix  $\mathbf{P}$  is called a regular matrix, when for some  $n$ , all entries of  $\mathbf{P}^n$  will be greater than zero, that is for some  $n$ ,  $P_{ij}^n > 0$ .

Consider the matrix:

$$\mathbf{P} = \begin{pmatrix} 0.2 & 0 & 0.8 \\ 0.5 & 0 & 0.5 \\ 0.3 & 0.7 & 0 \end{pmatrix}$$

Then

$$\mathbf{P}^2 = \begin{pmatrix} 0.28 & 0.56 & 0.16 \\ 0.25 & 0.35 & 0.4 \\ 0.41 & 0 & 0.59 \end{pmatrix} \text{ and } \mathbf{P}^3 = \begin{pmatrix} 0.384 & 0.112 & 0.504 \\ 0.345 & 0.280 & 0.375 \\ 0.259 & 0.413 & 0.328 \end{pmatrix}$$

Note that although the matrix  $\mathbf{P}$  has zero entries ( $P_{12} = P_{22} = P_{33} = 0$ ), in  $\mathbf{P}^3$  all entries are greater than zero, thus matrix  $\mathbf{P}$  is a regular matrix. A regular matrix will have stationary distribution and satisfy the system of equations in Eqs. (16.20) and (16.21).

#### EXAMPLE 16.4

The number of flights cancelled by an airline daily is modelled using a Markov chain. The states of the chain and the description of states are given in Table 16.8.

**TABLE 16.8** States representing cancellation of flights

State	Description
0	No cancellations
1	One cancellation
2	Two cancellations
3	More than 2 cancellations

The revenue loss (in millions of rupees) due to cancellation of flights in various states is given in Table 16.9.

**TABLE 16.9** Revenue loss due to cancellations

State	0	1	2	3
Loss	0	4.5	10.0	16.0

The transition probability matrix between states is shown in Table 16.10.

**TABLE 16.10** State transition matrix between flight cancellations

	0	1	2	3
0	0.45	0.30	0.20	0.05
1	0.15	0.60	0.15	0.10
2	0.10	0.30	0.40	0.20
3	0	0.10	0.70	0.20

- (a) If there are no cancellations initially, what is the probability that there will be at least one cancellation after 2 days?
- (b) Calculate the steady-state expected loss due to cancellation of flights.

### Solution:

- (a) If there are no cancellations initially, then the initial state vector is  $P_I = [1 \ 0 \ 0 \ 0]$ . The probability distribution after two days is

$$P_I P^2 = (1 \ 0 \ 0 \ 0) \begin{pmatrix} 0.45 & 0.30 & 0.20 & 0.05 \\ 0.15 & 0.60 & 0.15 & 0.10 \\ 0.10 & 0.30 & 0.40 & 0.20 \\ 0 & 0.10 & 0.70 & 0.20 \end{pmatrix}^2 = (0.2675 \ 0.38 \ 0.25 \ 0.1025)$$

Probability that there will be at least one cancellation after 2 days =  $0.38 + 0.25 + 0.1025 = 0.7325$ .

(b) To calculate the steady-state expected loss, we have to calculate the steady-state distribution. The steady-state distribution will satisfy the following system of equations:

$$\begin{aligned}\pi_0 &= 0.45\pi_0 + 0.15\pi_1 + 0.10\pi_2 \\ \pi_1 &= 0.30\pi_0 + 0.60\pi_1 + 0.30\pi_2 + 0.10\pi_3 \\ \pi_2 &= 0.20\pi_0 + 0.15\pi_1 + 0.40\pi_2 + 0.70\pi_3 \\ \pi_3 &= 0.05\pi_0 + 0.10\pi_1 + 0.20\pi_2 + 0.20\pi_3 \\ \pi_0 + \pi_1 + \pi_2 + \pi_3 &= 1\end{aligned}$$

Solving the above system of equations we get

$$(\pi_0 \quad \pi_1 \quad \pi_2 \quad \pi_3) = (0.163 \quad 0.390 \quad 0.311 \quad 0.137)$$

The steady-state expected loss is  $\sum_{i=0}^3 \pi_i \times L_i$ , where  $L_i$  is the expected revenue loss in state  $i$  (Table 16.9). Hence

$$\sum_{i=0}^3 \pi_i \times L_i = 0.163 \times 0 + 0.390 \times 4.5 + 0.311 \times 10 + 0.137 \times 16 = 7.05$$

## 16.5 | CLASSIFICATION OF STATES IN A MARKOV CHAIN

Not all Markov chains will have stationary probability distribution. To derive the necessary and sufficient conditions for existence of stationary distribution of a Markov chain, we have to understand different classes of states that exist in a Markov chain. In this section, we will be discussing classification of states and the necessary and sufficient conditions for existence of stationary distribution.

### 16.5.1 | Accessible State

A state  $j$  is accessible (or reachable) from state  $i$  if there exists a  $n$  such that  $P_{ij}^n > 0$ . That is, there exists a path from state  $i$  to state  $j$ .

### 16.5.2 | Communicating States

Two states  $i$  and  $j$  are communicating states when there exists  $n$  and  $m$  such that  $P_{ij}^n > 0$  and  $P_{ji}^m > 0$ . That is, state  $j$  can be reached (accessible) from state  $i$  and similarly state  $i$  can be reached from state  $j$ . A Markov chain is called **irreducible** if all states of the chain communicate with each other.

### 16.5.3 | Recurrent and Transient States

A state  $i$  of a Markov chain is called a recurrent state when

$$\sum_{n=1}^{\infty} P_{ii}^n = \infty \quad (16.27)$$

That is, if the state  $i$  is recurrent then the Markov chain will visit state  $i$  infinite number of times in the long run. If state  $i$  is recurrent and states  $i$  and  $j$  are communicating states, then state  $j$  is also a recurrent state.

A state  $k$  of a Markov chain is called a transient state when

$$\sum_{n=1}^{\infty} P_{kk}^n < \infty \quad (16.28)$$

That is, state  $k$  is called a transient state when  $\sum_{n=1}^{\infty} P_{kk}^n$  is finite. This means it is possible that the Markov chain may not return to state  $k$  in the long run.

#### 16.5.4 | First Passage Time and Mean Recurrence Time

First passage time is the probability that the Markov chain will enter state  $i$  exactly after  $n$  steps for the first time after leaving state  $i$ , that is

$$f_{ii}^n = P[X_n = i, X_k \neq i, k = 1, 2, \dots, n-1 | X_0 = i] \quad (16.29)$$

Let  $F_{ii} = \sum_{n=1}^{\infty} f_{ii}^n$ . Then for recurrent state  $F_{ii} = \sum_{n=1}^{\infty} f_{ii}^n = 1$  and for a transient state  $F_{ii} = \sum_{n=1}^{\infty} f_{ii}^n < 1$ . Mean recurrence time is the average time taken to return to state  $i$  after leaving state  $i$ . Mean recurrence time  $\mu_{ii}$  is given by

$$\mu_{ii} = \sum_{n=1}^{\infty} n \times f_{ii}^n \quad (16.30)$$

If the mean recurrence time is finite ( $\mu_{ii}$  is finite), then the recurrent state is called a **positive recurrent state** and if it is infinite then it is called **null-recurrent state**.

#### 16.5.5 | Periodic State

Periodic state is a special case of recurrent state in which  $d(i)$  is the greatest common divisor of  $n$  such that  $P_{ii}^n > 0$ . If  $d(i) = 1$ , it is called **aperiodic state** and if  $d(i) \geq 2$ , then it is called a **periodic state**. For example, consider a Markov chain shown in Table 16.11.

TABLE 16.11 Transition matrix

	1	2	3
1	0	1	0
2	0	0	1
3	1	0	0

In the matrix shown in Table 16.9, for state 1,  $P_{11}^3 = 1, P_{11}^6 = 1, P_{11}^9 = 1$ , and  $P_{11}^n = 0$  when  $n$  is not a multiple of 3. That is, the greatest common divisor is 3 which means that the periodicity is 3.

### 16.5.6 | Ergodic Markov Chain

A state  $i$  of a Markov chain is ergodic when it is positive recurrent and aperiodic. Markov chain in which all states are positive recurrent and aperiodic is called an **ergodic Markov chain**. For an ergodic Markov chain, a stationary distribution exists that satisfies the system of equations (16.20) and (16.21):

$$\begin{aligned}\pi_j &= \sum_{k=1}^m \pi_k P_{kj} \\ \sum_{k=1}^m \pi_k &= 1\end{aligned}$$

### 16.5.7 | Limiting Probability

In a Markov chain, the limiting probability is given by

$$\lim_{n \rightarrow \infty} p_{ij}^n$$

The main difference between limiting probability and stationary distribution is that stationary distribution when exists is unique and is not dependent on the initial state, whereas limiting probability may not be unique and may depend on the initial state.

## 16.6 | MARKOV CHAINS WITH ABSORBING STATES

A state  $i$  of a Markov chain is called an **absorbing state** when  $P_{ii} = 1$ , that is if the system enters this state, it will remain in the same state. Many real-life problems such as bad debt (non-performing assets), bankruptcy, customer churn, employee attrition and so on can be modelled using absorbing state Markov chain. Absorbing state Markov chain is a Markov chain in which there is at least one state  $k$  such that  $P_{kk} = 1$ . Non-absorbing states in an absorbing state Markov chain are transient states. Note that, an absorbing state Markov chain is not ergodic since states other than absorbing states will be transient states (that is,  $\sum_{n=1}^{\infty} P_{kk}^n < \infty$ ). The transition matrix corresponding to an absorbing state Markov chain is not a regular matrix and thus do not have stationary distribution. That is,  $P_i \times P^n$  may not converge to a unique value and will depend on the initial distribution. If there is a path connecting a transient state to an absorbing state, then it will be eventually absorbed. That is, in the long run the probability of finding the system in transient states will be zero. Absorbing state Markov chains have several practical applications. While using absorbing state Markov chains in analytics problem solving, we would like to learn the following from an absorbing state Markov chain:

1. The probability of eventual absorption to a specific absorbing state (when there are more than one absorbing states) from various transient states of the Markov chain.
2. The expected time to absorption from a transient state to absorbing states.

The above questions are answered using canonical form of the transition matrix.

### 16.6.1 | Canonical Form of the Transition Matrix of an Absorbing State Markov Chain

The rows of the transition probability matrix of an absorbing state Markov chain can be rearranged such that the top rows are assigned for absorbing states followed by transient states (the idea here is to group the absorbing state and non-absorbing states). Let  $A$  and  $T$  be the set of absorbing and transient states, respectively, in the Markov chain. Then the transition probability matrix can be arranged such that

$$\mathbf{P} = \begin{pmatrix} & A & T \\ A & I & 0 \\ T & R & Q \end{pmatrix}$$

The matrix  $\mathbf{P}$  is divided into 4 matrices  $I$ ,  $0$ ,  $R$ , and  $Q$ , where

1. Matrix  $I$  is the identity matrix. It corresponds to transition within absorbing states.
2. Matrix  $0$  is a matrix in which all elements are zero. Here the elements correspond to transition from an absorbing state to transient states.
3. Matrix  $R$  represents the probability of absorption from a transient state to an absorbing state.
4. Matrix  $Q$  represents the transition between transient states.

To calculate the eventual probability of absorption, we would like to calculate the long-run (limiting probability) value of  $R$  in the above matrix. When we multiply the canonical form of the matrix, we get

$$\mathbf{P}^n = \begin{pmatrix} I & 0 \\ \left( \sum_{k=0}^{n-1} Q^k \right) R & Q^n \end{pmatrix} \quad (16.31)$$

For large  $n$ , the matrix  $\left( \sum_{k=0}^{n-1} Q^k \right) R$  will give the probability of eventual absorption to an absorbing state.

As  $n \rightarrow \infty$ , we can show that  $\sum_{k=0}^{n-1} Q^k = F = (I - Q)^{-1}$ . The matrix  $F$  is called the fundamental matrix and the matrix  $FR$  is the probability of eventual absorption into an absorbing state from a transient state. The expected time to absorption is given by

$$\text{Expected time to absorption} = Fc \quad (16.32)$$

where  $c$  is a unit vector. That is, the row sum of the fundamental matrix gives the expected duration for absorption (that is, expected time it takes to reach an absorbing state from a transient state).

#### EXAMPLE 16.5

Airwaves India (AI) is a mobile phone service provider based in Allahabad, India that provides several value-added services such as mobile data, video conferencing, etc. The market is highly competitive and AI faces high churn rate among its customers. The customers of AI are categorized into different states as listed below:

**STATE 1**

Customer churn that generated no revenue/profit

**STATE 2**

Customer churn that generated INR 200 profit per month on average (customer uses the service only for incoming calls and data)

**STATE 3**

Customer state that generated INR 300 profit per month on average

**STATE 4**

Customer state that generated INR 400 profit per month on average

**STATE 5**

Customer state that generated INR 600 profit per month on average

**STATE 6**

Customer state that generated INR 800 profit per month on average

The transition probability values between different states are shown in Table 16.12.

**TABLE 16.12** Transition probability matrix (based on monthly data)

	1	2	3	4	5	6
1	1	0	0	0	0	0
2	0	1	0	0	0	0
3	0.05	0.05	0.90	0	0	0
4	0.10	0.05	0	0.80	0.05	0
5	0.20	0.10	0	0.05	0.60	0.05
6	0.10	0.20	0	0	0	0.70

- (a) If a customer is in state 6, calculate the probability of eventual absorption in state 2?
- (b) Calculate the expected value of time taken to absorption if the current state is 4.

**Solution:**

- (a) To calculate the probability of absorption of a customer in state 6 to state 2, we have to calculate **FR**.

The matrix  $\mathbf{Q}$  is given by

$$\mathbf{Q} = \begin{bmatrix} 0.9 & 0 & 0 & 0 \\ 0 & 0.8 & 0.05 & 0 \\ 0 & 0.05 & 0.6 & 0.05 \\ 0 & 0 & 0 & 0.7 \end{bmatrix}$$

$$\mathbf{I} - \mathbf{Q} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0.9 & 0 & 0 & 0 \\ 0 & 0.8 & 0.05 & 0 \\ 0 & 0.05 & 0.6 & 0.05 \\ 0 & 0 & 0 & 0.7 \end{bmatrix} = \begin{bmatrix} 0.1 & 0 & 0 & 0 \\ 0 & 0.2 & -0.05 & 0 \\ 0 & -0.05 & 0.4 & -0.05 \\ 0 & 0 & 0 & 0.3 \end{bmatrix}$$

$$\mathbf{F} = (\mathbf{I} - \mathbf{Q})^{-1} = \begin{bmatrix} 10 & 0 & 0 & 0 \\ 0 & 5.1613 & 0.6452 & 0.1075 \\ 0 & 0.6452 & 2.5806 & 0.4301 \\ 0 & 0 & 0 & 3.3333 \end{bmatrix}$$

Probability of absorption  $\mathbf{FR}$  is given by

$$\mathbf{FR} = \begin{bmatrix} 10 & 0 & 0 & 0 \\ 0 & 5.1613 & 0.6452 & 0.1075 \\ 0 & 0.6452 & 2.5806 & 0.4301 \\ 0 & 0 & 0 & 3.3333 \end{bmatrix} \times \begin{bmatrix} 0.05 & 0.05 \\ 0.1 & 0.05 \\ 0.2 & 0.1 \\ 0.1 & 0.2 \end{bmatrix} = \begin{bmatrix} 0.5 & 0.5 \\ 0.6559 & 0.3441 \\ 0.6237 & 0.3763 \\ 0.3333 & 0.6667 \end{bmatrix}$$

That is, if the current customer state is 6, the probability of absorption into churn state 2 is 0.6667.

(b) Expected value of time to absorption is given by  $\mathbf{Fc}$ :

$$\mathbf{Fc} = \begin{bmatrix} 10 & 0 & 0 & 0 \\ 0 & 5.1613 & 0.6452 & 0.1075 \\ 0 & 0.6452 & 2.5806 & 0.4301 \\ 0 & 0 & 0 & 3.3333 \end{bmatrix} \times \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 10 \\ 5.91 \\ 3.65 \\ 3.33 \end{bmatrix}$$

Expected value of time to absorption when the current state is 4 is 5.91 months.

## 16.7 | EXPECTED DURATION TO REACH A STATE FROM OTHER STATES

Expected duration to reach a state of a Markov chain from another state of the chain is important when there is an undesirable state such as proportion of non-performing assets in a bank or low credit rating, etc. In such cases, the duration to reach such states from other states can be used as an early warning system. The expected duration to reach a state  $j$  from state  $i$  can be calculated by solving the following

system of equations: Let  $E_{i,j}$  = Expected number of duration to reach state  $j$  from state  $i$ . Then,  $E_{i,j}$  satisfies the following system of equations:

$$E_{i,j} = 1 + \sum_k P_{i,k} E_{k,j} \quad \forall i, i \neq j \quad (16.33)$$

$$E_{jj} = 0 \quad (16.34)$$

### EXAMPLE 16.6

The percentage of non-performing assets at a bank is classified into the following seven states:

State	State Description
1	NPA is less than 1%
2	NPA is between 1% and 2%
3	NPA is between 2% and 3%
4	NPA is between 3% and 4%
5	NPA is between 4% and 5%
6	NPA is between 5% and 6%
7	NPA greater than 6%

The transition probability matrix calculated based on monthly data is shown in Table 16.13.

TABLE 16.13 Transition probability matrix between NPA states

	1	2	3	4	5	6	7
1	0.95	0.05	0	0	0	0	0
2	0.10	0.85	0.05	0	0	0	0
3	0	0.10	0.80	0.10	0	0	0
4	0	0	0.15	0.70	0.15	0	0
5	0	0	0	0.15	0.65	0.20	0
6	0	0	0	0	0.20	0.60	0.20
7	0	0	0	0	0	0.10	0.90

Calculate the expected duration (in months) for the process to reach state 7 from state 4.

#### Solution:

Let  $E_{4,7}$  be the expected number of duration for the process to reach state 7 from state 4. Then it satisfies the following system of equations:

$$E_{4,7} = 1 + 0.15E_{3,7} + 0.70E_{4,7} + 0.15E_{5,7}$$

$$\begin{aligned}
E_{3,7} &= 1 + 0.10E_{2,7} + 0.80E_{3,7} + 0.10E_{4,7} \\
E_{5,7} &= 1 + 0.15E_{4,7} + 0.65E_{5,7} + 0.20E_{6,7} \\
E_{6,7} &= 1 + 0.20E_{5,7} + 0.60E_{6,7} + 0.20E_{7,7} \\
E_{2,7} &= 1 + 0.10E_{1,7} + 0.85E_{2,7} + 0.05E_{3,7} \\
E_{1,7} &= 1 + 0.95E_{1,7} + 0.05E_{2,7} \\
E_{7,7} &= 0
\end{aligned}$$

Solving the system of equations we get  $E_{4,7} = 206.6667$ . That is, it takes approximately 207 months on average for the process to reach state 7 from state 4.

## 16.8 | CALCULATION OF RETENTION PROBABILITY AND CUSTOMER LIFETIME VALUE USING MARKOV CHAINS

One of the important applications of Markov chains in marketing and customer relationship management (CRM) is calculation of retention probability and customer lifetime value (CLV). CLV is the net present value (NPV) of the future margin generated from a customer or a customer segment. CLV is calculated usually at a customer segment level. The customer segments can be represented as states of the Markov chain. Let  $\{0, 1, 2, \dots, m\}$  be the states of a Markov chain in which states  $\{1, 2, \dots, m\}$  denote different customer segments and state 0 denotes non-customer state. Ching *et al.* (2004) showed that the steady-state retention probability can be calculated using

$$R_i = \sum_{i=1}^m \frac{\pi_i}{\left( \sum_{j=1}^m \pi_j \right)} (1 - P_{i0}) = 1 - \frac{\pi_0 (1 - P_{00})}{1 - \pi_0} \quad (16.35)$$

where  $R_i$  is the steady-state retention probability.  $\sum_{i=1}^m \frac{\pi_i}{\left( \sum_{j=1}^m \pi_j \right)} (1 - P_{i0})$  is probability of retaining the

customer as customer (not necessarily in the same state) in the steady state:

$$\sum_{i=1}^m \frac{\pi_i}{\left( \sum_{j=1}^m \pi_j \right)} (1 - P_{i0}) = \sum_{i=1}^m \frac{\pi_i}{\left( \sum_{j=1}^m \pi_j \right)} - \sum_{i=1}^m \frac{\pi_i}{1 - \pi_0} \times P_{i0} = 1 - \frac{\pi_0 (1 - P_{00})}{1 - \pi_0}$$

using the relationship  $\sum_{i=1}^m \pi_i \times P_{i0} = \pi_0 (1 - P_{00})$ .

The customer lifetime value for  $N$  periods is given by (Pfeifer and Carraway, 2000):

$$\text{CLV} = \sum_{t=0}^N \frac{\mathbf{P}_I \times \mathbf{P}^t \mathbf{R}}{(1+i)^t} \quad (19.36)$$

where  $\mathbf{P}_I$  is the initial distribution of customers in different states,  $\mathbf{P}$  is the transition probability matrix,  $\mathbf{R}$  is the reward vector (margin generated in each customer segment). The interest rate is  $i$  (discount rate),  $d = 1/(1+i)$  is the discount factor.

### EXAMPLE 16.7

The customers of Dubai Data Services (DDS) are classified into five categories as shown in Table 16.14 along with transition probability matrix. State 0 represents non-customers and the remaining states are different customer segments created based on the revenue generated. The average margin generated in different states is shown in Table 16.15 along with initial distribution of customers in millions. Calculate the steady-state retention probability and CLV for 6 periods ( $N = 5$ ) using a discount factor of  $d = 0.95$ .

**TABLE 16.14** Customer states of DDS and transition matrix

	0	1	2	3	4
0	0.80	0.10	0.10	0	0
1	0.10	0.60	0.20	0.10	0
2	0.15	0.05	0.75	0.05	0
3	0.20	0	0.10	0.60	0.10
4	0.30	0	0	0.05	0.65

**TABLE 16.15** Margin generated in different states

State	0	1	2	3	4
Average Margin	0	120	300	450	620
Customers (in millions)	55.8	6.5	4.1	2.3	1.6

#### Solution:

The stationary distribution equations are

$$\begin{aligned}\pi_0 &= 0.8\pi_0 + 0.10\pi_1 + 0.15\pi_2 + 0.20\pi_3 + 0.30\pi_4 \\ \pi_1 &= 0.1\pi_0 + 0.60\pi_1 + 0.05\pi_2 \\ \pi_2 &= 0.1\pi_0 + 0.2\pi_1 + 0.75\pi_2 + 0.10\pi_3 \\ \pi_3 &= 0.10\pi_1 + 0.05\pi_2 + 0.60\pi_3 + 0.05\pi_4 \\ \pi_0 + \pi_1 + \pi_2 + \pi_3 + \pi_4 &= 1\end{aligned}$$

Solving the above system of equations, we get  $\pi_0 = 0.4287$ . The steady-state retention probability  $R_t$  is

$$R_t = 1 - \frac{\pi_0(1-P_{00})}{1-\pi_0} = 1 - \frac{0.4287 \times (1-0.80)}{1-0.4287} = 0.85$$

Customer lifetime value for  $N = 5$  is

$$\text{CLV} = \sum_{t=0}^5 \frac{\mathbf{P}_I \times \mathbf{P}^t \mathbf{R}}{(1+i)^t}$$

where

$$\mathbf{P}_I = \begin{pmatrix} 55.8 & 6.5 & 4.1 & 2.3 & 1.6 \end{pmatrix}$$

$$\mathbf{R} = \begin{pmatrix} 0 \\ 120 \\ 300 \\ 450 \\ 620 \end{pmatrix}$$

Substituting the values in CLV equation, we get  $\text{CLV} = 40181.59$ .

## 16.9 | MARKOV DECISION PROCESS (MDP)

Markov decision process (MDP) is a powerful technique based on Markov chain that can be used for analysing any sequential decision-making scenario over a planning horizon. For example, consider an investor in stock market. He has to decide whether to buy or sell stocks during each investment period. Another example of sequential decision making is a football match. The coach has to decide when to make substitutions which can have significant impact on the outcome of the match. In this case, the planning horizon is 90 minutes and the duration of 90 minutes can be broken into several smaller intervals in which the coach has to take decision whether to make a substitution or not. First successful business application of MDP was reported by Ronald Howard (2002) in which MDP was used for catalog mailing policy (whether to send a catalog or not to a customer during a season) for the American retail chain Sears. Following are few examples of sequential decision-making scenarios which can be analysed using MDP:

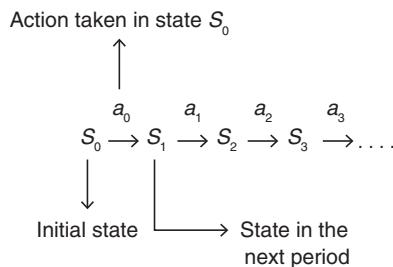
1. Decision on whether to promote a product and what promotion strategy to use.
2. When to change job, car, and any capital equipment (such as aircraft).
3. When to buy and sell shares.
4. Movement of robots in a given context (they have to decide next action based on the current state).
5. When to stop (or change) a television serial with an objective to maximize television rating points (TRP).

MDP is defined using a 5-tuple  $(\mathbf{S}, \mathbf{A}, \mathbf{P}_{sa}, \mathbf{R}, \beta)$  where

1.  $\mathbf{S}$  is the state space where  $\mathbf{S} = \{S_1, S_2, \dots, S_n\}$
2.  $\mathbf{A}$  is the set of actions  $\mathbf{A} = \{a_1, a_2, \dots, a_m\}$ . Not all actions will be available for all states. For example, if MDP is used for buying and selling shares, the action 'sell' will be available to the decision maker only when he/she owns a stock.

3.  $P_{sa}$  are state transition probability values that depend on the current state and the chosen action.
4.  $R$  is the reward vector that is a mapping of  $S \times A \rightarrow R$ . The reward depends on the state and the action chosen.
5.  $\beta$  is the discount factor,  $\beta \in [0, 1]$ . Discount factor is used since the rewards are received at future times.  $\beta = \frac{1}{1+i}$ , where  $i$  is the interest rate.

In a Markov decision process, the decision maker takes series of decisions based on the state that results in immediate reward and future reward. MDP is a reinforcement learning algorithm since there is uncertainty about the rewards that are likely to occur at a future time points. The objective is to maximize the expected total reward obtained over a period of time. Figure 16.4 shows the sequential decisions associated with a MDP.



**FIGURE 16.4** Illustration of Markov decision process.

Based on the state and action, a reward is generated. MDP will generate a sequence of reward as shown in Eq. (16.37).

$$R(S_0, a_0) + \beta R(S_1, a_1) + \beta^2 R(S_2, a_2) + \beta^3 R(S_3, a_3) + \dots \quad (16.37)$$

In Eq. (16.37),  $R(S_0, a_0)$  is the reward generated from the initial state  $S_0$  and action  $a_0$  is taken by the decision maker. The reward  $R(S_1, a_1)$  is discounted by  $\beta$  since this is a reward received at a future time. The objective of MDP is to find the optimal action sequence  $(a_0, a_1, a_2 \dots)$  that will maximize the total expected reward. That is

$$\underset{a_i \in A}{\text{Maximize}} E[R(s_0, a_0) + \beta R(s_1, a_1) + \beta^2 R(s_2, a_2) + \dots] \quad (16.38)$$

MDP is solved using the following two algorithms:

1. Policy iteration algorithm
2. Value iteration algorithm

Policy iteration finds the optimal policy which is a mapping from state space to action space that maximizes the NPV of total reward based on infinite planning horizon. On the other hand, the value iteration algorithm uses dynamic programming to find the optimal actions for a finite planning horizon.

### 16.9.1 | Policy Iteration Algorithm

A policy  $\Pi$  in a MDP is a function that maps states to actions, that is,  $\Pi: S \rightarrow A$ . For example, assume a Markov chain with four states  $\{S_1, S_2, S_3, S_4\}$  and assume that the action set  $A$  has three possible actions, that is  $A = \{a_1, a_2, a_3\}$ . Then policy  $\Pi = \{a_1, a_1, a_2, a_3\}$  implies that whenever the state of the system is  $S_1$  or  $S_2$ , action  $a_1$  is chosen; when the state is  $S_3$ , action  $a_2$  is chosen; and when the state is  $S_4$ , action  $a_3$  is chosen. The objective of the policy iteration algorithm is to find an optimal policy that will maximize the value function of the Markov decision process (NPV of the total reward). The value function of a policy  $\Pi$  is the expected sum of discounted rewards starting from state  $S_i$  and following the policy  $\Pi$ . The value function for policy  $\Pi$  stating at state  $S_i$  is given by

$$V^\Pi(i) = \underbrace{R^\Pi(i)}_{\text{Immediate reward}} + \beta \underbrace{\sum_{j \in S} P_{ij}^\Pi \times V^\Pi(j)}_{\text{Discounted future reward}} \quad (16.39)$$

where  $V^\Pi(i)$  is the value function for policy  $\Pi$  when the initial state is  $S_i$ ;  $R^\Pi(i)$  is the initial reward when action  $a_i$  is taken when the state is  $S_i$  according to policy  $\Pi$ ;

$P_{ij}^\Pi$  is the transition probability from state  $i$  to state  $j$  when action  $a_i$  is taken when the initial state is  $S_i$  according to policy  $\Pi$ .

The policy iteration algorithm has two steps (Howard, 1971): (a) Policy evaluation and (b) policy improvement.

**Policy Evaluation Step:** Let  $S = \{S_1, S_2, \dots, S_p, \dots, S_n\}$  be the set of states and  $A = \{a_1, a_2, \dots, a_m\}$  be the set of actions. Let  $\Pi$  be a policy which maps the state space to the action space and the decision maker always chooses the action based on the policy  $\Pi$ . Write the system of the value function  $V^\Pi(i)$  equations for states  $\{S_1, S_2, \dots, S_p, \dots, S_n\}$ . That is

$$V^\Pi(i) = R^\Pi(i) + \beta \sum_{j=S_1}^{S_n} P_{ij}^\Pi V^\Pi(j) \quad \forall i \in S \quad (16.40)$$

The system of linear equations in Eq. (16.40) can be solved to find the value function  $V^\Pi(i)$  for all states of the MDP.

**Policy Improvement Step:** In the policy improvement step, the objective is to improve the value function obtained by solving the system of equations in Eq. (16.40) by finding a better policy if it exists. Let  $\Pi = \{a_1, a_2, \dots, a_p, \dots, a_n\}$  be the current policy, that is for state  $S_i$  the action in the current policy is  $a_i$ . In the policy improvement, the following steps are used:

#### STEP 1

For  $i \in S$ , compute

$$T^{\Pi^{\text{new}}}(i) = \max_{a_{i,\text{new}}} \left( R(i, a_{i,\text{new}}) + \beta \sum_{j=S_1}^{S_n} P(j|i, a_{i,\text{new}}) V^\Pi(j) \right) \quad (16.41)$$

where

$a_{i,\text{new}}$  = A new action chosen for state  $i$ .

$T^{\Pi^{\text{new}}}(i)$  = Value function when the current policy for state  $i$  is changed to  $a_{i,\text{new}}$ .

$R(i, a_{i,\text{new}})$  = Reward when action for state  $i$  is replaced with new action  $a_{i,\text{new}}$ .

## STEP 2

If  $T^{\Pi^{\text{new}}}(i) > V^\Pi(i)$ , then replace the policy  $\Pi$  with the new policy  $\Pi^{\text{new}}$  (that is, action  $a_i$  in policy  $\Pi$  is replaced with  $a_{i,\text{new}}$ ) else retain policy  $\Pi$ . Repeat steps 1 and 2 for all states and actions.

### EXAMPLE 16.8

States of a mining equipment used in coal mining are classified into 4 states {1, 2, 3, 4}. States represent the condition of the mining equipment, 1 being excellent condition and 4 being bad condition. The maintenance department may decide to take the following actions:

1. Do nothing
2. Carry out a preventive maintenance which changes states 3 and 4 to 2 and thus carried out only when the state of the system is either 3 or 4. The cost of preventive maintenance is INR 2000.
3. Replace the equipment with new equipment (that will change the state of equipment to state 1). The cost of replacement is INR 10,000. Replacement can be used when the equipment is in states 2, 3, and 4

The mining equipment generates revenue of INR 20,000 in state 1, INR 16,000 in state 2, INR 12,000 in state 3, and INR 5,000 in state 4 during each period.

The state transitions between 4 states are shown in Table 16.16.

TABLE 16.16 Transition probability matrix

	1	2	3	4
1	0.8	0.1	0.1	0
2	0	0.7	0.2	0.1
3	0	0	0.7	0.3
4	0	0	0	1

Calculate the value function for the policy {1, 1, 2, 2} and check whether policy {1, 1, 2, 3} is better than policy {1, 1, 2, 2}. Use a discount factor  $\beta = 0.95$ . Assume that

the actions are taken at the beginning of each period and preventive maintenance and replacement times are negligible.

### Solution:

Let  $\Pi = \{1, 1, 2, 2\}$ . The corresponding value function for different states  $\{1, 2, 3, 4\}$  is given by (policy evaluation step)

$$\begin{aligned} V^\Pi(1) &= 20,000 + 0.95[0.8V^\Pi(1) + 0.1V^\Pi(2) + 0.1V^\Pi(3)] \\ V^\Pi(2) &= 16,000 + 0.95[0.7V^\Pi(2) + 0.2V^\Pi(3) + 0.1V^\Pi(4)] \\ V^\Pi(3) &= 14,000 + 0.95[0.7V^\Pi(2) + 0.2V^\Pi(3) + 0.1V^\Pi(4)] \\ V^\Pi(4) &= 14,000 + 0.95[0.7V^\Pi(2) + 0.2V^\Pi(3) + 0.1V^\Pi(4)] \end{aligned}$$

In the above system of equations, the action for state 3 is 2 (preventive maintenance), cost of preventive maintenance is 2000, and the revenue generated in state 2 is 16000, so the net immediate reward is  $16000 - 2000 = 14000$ .

Solving the system of equations we get

$$V^\Pi(1) = 326850, V^\Pi(2) = 308600, V^\Pi(3) = 306600, \text{ and } V^\Pi(4) = 306600$$

The total discounted reward is  $\sum_{i=1}^4 V^\Pi(i) = 1248650$

**Policy Improvement Step:** Consider the policy  $\{1, 1, 2, 3\}$ . Under this policy the action for state 4 is changed to action 3 (replacement) from 2 (preventive maintenance). The corresponding policy improvement equation is

$$T^{\Pi^{new}} = 10,000 + 0.95[0.8V^\Pi(1) + 0.1V^\Pi(2) + 0.1V^\Pi(3)] \quad (16.42)$$

In the above equation, the new action for state 4 is 3 (replacement at the cost of 10,000) and the new state will be 1 (in which a revenue of 20,000 is generated). The net immediate reward is  $20,000 - 10,000 = 10,000$ . Substituting the values of  $V^\Pi(1)$ ,  $V^\Pi(2)$ , and  $V^\Pi(3)$  in the Eq. (16.42), we get

$$T^{\Pi^{new}} = 10,000 + 0.95[0.8 \times 326850 + 0.1 \times 308600 + 0.1 \times 306600] = 316850$$

Since  $T^{\Pi^{new}}(4) > V^\Pi(4)$ , we can conclude that the new policy  $\{1, 1, 2, 3\}$  is better than the policy  $\{1, 1, 2, 2\}$ . Solving the policy evaluation for the new policy  $\{1, 2, 2, 3\}$  we get

$$V^{\Pi^{new}}(1) = 337895.5; V^{\Pi^{new}}(2) = 322552.2; V^{\Pi^{new}}(3) = 320552.2; V^{\Pi^{new}}(4) = 327895.5$$

$$\sum_{i=1}^4 V^{\Pi^{new}}(i) = 1308896$$

The new policy  $\{1, 1, 2, 3\}$  is better than the policy  $\{1, 1, 2, 2\}$ , but not necessarily optimal. We have to repeat the policy improvement step to arrive at an optimal solution. Alternatively, we can use linear programming formulation to find the optimal policy.

### 16.9.2 | Linear Programming Formulation for Finding Optimal Policy

The optimal policy can be calculated by formulating the problem as a linear programming problem. Let  $\Pi^*$  be the optimal policy for the MDP. Then  $\Pi^*$  should satisfy the following constraints:

$$V^{\Pi^*}(1) \geq 20000 + 0.95[0.8V^{\Pi^*}(1) + 0.1V^{\Pi^*}(2) + 0.1V^{\Pi^*}(3)]$$

In state 1, only action 1 will be used since other actions do not make sense.

Constraint for state 2 and action 1 is

$$V^{\Pi^*}(2) \geq 16000 + 0.95[0.7V^{\Pi^*}(2) + 0.2V^{\Pi^*}(3) + 0.1V^{\Pi^*}(4)]$$

Constraint for state 2 and action 3 is

$$V^{\Pi^*}(2) \geq 10000 + 0.95[0.8V^{\Pi^*}(1) + 0.1V^{\Pi^*}(2) + 0.1V^{\Pi^*}(3)]$$

Constraint for state 3 and action 1 is

$$V^{\Pi^*}(3) \geq 12000 + 0.95[0.7V^{\Pi^*}(3) + 0.3V^{\Pi^*}(4)]$$

Constraint for state 3 and action 2 is

$$V^{\Pi^*}(3) \geq 14000 + 0.95[0.7V^{\Pi^*}(2) + 0.2V^{\Pi^*}(3) + 0.1V^{\Pi^*}(4)]$$

Constraint for state 3 and action 3 is

$$V^{\Pi^*}(3) \geq 10000 + 0.95[0.8V^{\Pi^*}(1) + 0.1V^{\Pi^*}(2) + 0.1V^{\Pi^*}(3)]$$

Constraint for state 4 and action 1 is

$$V^{\Pi^*}(4) \geq 5000 + 0.95V^{\Pi^*}(4)$$

Constraint for state 4 and action 2 is

$$V^{\Pi^*}(4) \geq 14000 + 0.95[0.7V^{\Pi^*}(2) + 0.2V^{\Pi^*}(3) + 0.1V^{\Pi^*}(4)]$$

Constraint for state 4 and action 3 is

$$V^{\Pi^*}(4) \geq 10000 + 0.95[0.8V^{\Pi^*}(1) + 0.1V^{\Pi^*}(2) + 0.1V^{\Pi^*}(3)]$$

The objective function is

$$\text{Minimize } \sum_{i=1}^4 V^{\Pi^*}(i)$$

Note that the objective function in this case is minimization. This is due to the fact that when the objective function is maximization and if all the constraints are greater than or equal to constraints then the feasible region will be unbounded and we cannot solve the problem. The binding constraint for each state forms the optimal policy. The Excel Solver solution for the above formulation is shown in Table 16.17.

**TABLE 16.17** Excel solver output for LP formulation

$V^{\Pi^*}(1)$	$V^{\Pi^*}(2)$	$V^{\Pi^*}(3)$	$V^{\Pi^*}(4)$	Objective Function Value
362000	352000	352000	352000	1418000
Constraints				
LHS	RHS	(State, Action)	Constraint Type	
20000	20000	(1,1)	Binding	
17600	16000	(2,1)	Non-Binding	
10000	10000	(2,3)	Binding	
17600	12000	(3,1)	Non-Binding	
17600	14000	(3,2)	Non-Binding	
10000	10000	(3,3)	Binding	
17600	5000	(4,1)	Non-Binding	
17600	14000	(4,2)	Non-Binding	
10000	10000	(4,3)	Binding	

The optimal policy is  $\Pi^* = \{1, 3, 3, 3\}$  and the optimal values are  $V^{\Pi^*}(1) = 362000$ ,  $V^{\Pi^*}(2) = V^{\Pi^*}(3) = V^{\Pi^*}(4) = 352000$ . The total discounted reward is 1418000.

## 16.10 | VALUE ITERATION ALGORITHM

The policy iteration algorithm finds the optimal policy based on infinite planning horizon using a two-step procedure of policy evaluation and policy improvement. In many cases, we may have to find optimal policy when the planning horizon is finite; in such cases the value iteration algorithm is used. Value iteration algorithm is basically a dynamic programming algorithm which uses divide-and-conquer strategy to solve the problem. Dynamic programming is used when the problem fits into the following characteristics:

1. Problem can be divided into **stages**. In MDP, the stages are different time periods during which decisions need to be taken.
2. The system is identified in each stage using a **state**. In MDP, state is the information required to take decision and find the optimal solution up to that stage.
3. The **decision** taken at each stage updates the state at the next stage of the planning horizon.

Dynamic programming is based on Bellman's Principle of Optimality (Bellman, 1957) stated below:

An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to state resulting from the first decision.

Bellman's principle of optimality allows the problem to be broken into smaller problems which can be solved easily. The following steps are used in value iteration algorithm:

1. Identify the stages of the MDP (time periods).
2. Identify the state at the beginning of each stage (information needed for choosing action).
3. Identify the actions (decisions) available for the decision maker at each state.
4. Write the recursive relationship between the optimal action (decision) at current stage and the previous optimal action.
5. Optimal action set is the optimal policy for the problem up to that stage.

The dynamic programming recursive equation for the value iteration algorithm is given by

$$V_t^*(i) = \max_{a_i \in A} \left( R(i, a_i) + \beta \sum_{j=S_1}^{S_n} P_{ij}(a_i) V_{t+1}^*(j) \right) \quad (16.43)$$

where  $V_t^*(i)$  is the optimal value for a policy when the current period is  $t$  and the current state is  $S_i$ . Equation (16.43) is a backward recursive equation. If the duration of the planning horizon is  $n$ , then we assume that  $V_{n+1}^*(i) = 0$  for all states  $S_i$ .

### EXAMPLE 16.9

Find the optimal policy for Example 16.5 when the planning horizon is 4 periods.

**Solution:**

The total duration is 4, so  $V_5^* = 0$  for  $i = 1, 2, 3$ , and 4. Now

$$V_4^*(1) = \max_{a_1 \in A} \left( R(1, a_1) + \beta \sum_{j=1}^4 P_{j1}(a_1) V_5^*(j) \right)$$

The optimal action for state 1 when  $t = 4$  is 'do nothing' (action 1) and the corresponding value is

$$V_4^*(1) = \max_{a_1 \in A} (R(1, 1)) = 20000$$

The optimal action for state 2 when  $t = 4$  is 'do nothing' (action 1)

$$\begin{aligned} V_4^*(2) &= \max_{a_2 \in A} \left( R(2, a_2) + \beta \sum_{j=1}^4 P_{j2}(a_2) V_5^*(j) \right) = \max(R(2, 1); (2, 3)) \\ &= \max(16,000; 10000) = 16,000 \end{aligned}$$

The optimal action for state 3 when  $t = 4$  is 'do preventive maintenance' (action 2)

$$\begin{aligned} V_4^*(3) &= \max_{a_3 \in A} \left( R(3, a_3) + \beta \sum_{j=1}^4 P_{j3}(a_3) V_5^*(j) \right) = \max(R(3, 1); R(3, 2); R(3, 3)) \\ &= \max(12000; 14000; 10000) = 14000 \end{aligned}$$

The optimal action for state 4 when  $t = 4$  is 'do preventive maintenance' (action 2) and the corresponding recursive equation is

$$\begin{aligned} V_4^*(4) &= \max_{a_i \in A} \left( R(4, a_i) + \beta \sum_{j=1}^4 P_{ij}(a_i) V_5^*(j) \right) = \max(R(4,1); R(4,2); R(4,3)) \\ &= \max(5000; 14000; 10000) = 14000 \end{aligned}$$

The optimal policy when  $t = 4$  is (1, 1, 2, 2).

For  $t = 3$  and  $i = 1$ , the recursive equation is given by

$$V_3^*(1) = \max_{a_i \in A} \left( R(1, a_i) + \beta \sum_{j=1}^4 P_{ij}(a_i) V_4^*(j) \right)$$

$$V_3^*(1) = \max(R(1,1) + 0.95(0.8 * 20000 + 0.1 * 16000 + 0.1 * 14000)) = 38050$$

$$V_3^*(2) = \max_{a_i \in A} \left( R(2, a_i) + \beta \sum_{j=1}^4 P_{ij}(a_i) V_4^*(j) \right)$$

$$V_3^*(2) = \max \left( \begin{array}{l} R(2,1) + 0.95(0.7 * 16000 + 0.2 * 14000 + 0.1 * 14000); \\ R(2,3) + 0.95 (0.8 * 20000 + 0.1 * 16000 + 0.1 * 14000) \end{array} \right) = 30630$$

$$V_3^*(3) = \max_{a_i \in A} \left( R(3, a_i) + \beta \sum_{j=1}^4 P_{ij}(a_i) V_4^*(j) \right)$$

$$\begin{aligned} &= \max[R(3,1) + 0.95(0.7 * 14000 + 0.3 * 14000); R(3,2) \\ &\quad + 0.95(0.7 * 16000 + 0.2 * 14000 + 0.1 * 14000); R(3,3) \\ &\quad + 0.95 * (0.8 * 20000 + 0.1 * 16000 + 0.1 * 14000)] \end{aligned}$$

$$V_3^*(3) = \max(25300, 28630, 28050) = 28630$$

Optimal actions for state 3 for  $t = 3$  is actions 2 (perform preventive maintenance).

For state 4,  $t = 3$  the optimal value is

$$\begin{aligned} V_3^*(4) &= \max_{a_i \in A} \left( R(4, a_i) + \beta \sum_{j=1}^4 P_{ij}(a_i) V_4^*(j) \right) \\ &= \max [R(4, 1) + 0.95(14000); R(4,2) + 0.95 (0.7 * 16000 + 0.2 * 14000 \\ &\quad + 0.1 * 14000); R(4, 3) + 0.95 * (0.8 * 20000 + 0.1 * 16000 + 0.1 * 14000)] \end{aligned}$$

$$V_3^*(4) = \max(18300, 28630, 28050) = 28630$$

Optimal action for state 4 for  $t = 3$  is action 2 (perform preventive maintenance).

The above recursive relation can be repeated. The optimal policy for states and duration of planning horizon are shown in Table 16.18. The optimal policy when there are four periods in the planning horizon is (1, 3, 3, 3).

**TABLE 16.18** Value iteration algorithm

<i>t=1</i>		<i>t=2</i>		<i>t=3</i>		<i>t=4</i>	
State	Optimal Action						
1	1	1	1	1	1	1	1
2	3	2	3	2	1	2	1
3	3	3	3	3	2	3	2
4	3	4	3	4	2	4	2

The optimal values are given by

$$V_1^*(1) = 69410.317; V_1^*(2) = 59920.32; V_1^*(3) = 59920.32; V_1^*(4) = 59920.32$$

## SUMMARY

1. Most problems in analytics are dynamic in nature and thus require collection of random variables to model the problem.
2. Stochastic process is a collection of random variables usually indexed by time  $t$  and used while modelling problems that are not independent and identically distributed.
3. Poisson process is a counting process that is used in decision-making scenarios such as capacity planning and spare parts demand forecasting. Compound Poisson process can be used to study problems such as cash replenishments at ATMs, total insurance claims, etc.
4. Markov chain is one of the most powerful models in analytics with applications across industry sectors. Google's PageRank algorithm is based on Markov chain.
5. Asset availability, market share, customer retention probability, and customer lifetime value are few applications of Markov chain in analytics.
6. Absorbing state Markov chain can be used for modelling problems such as non-performing assets and customer churn.
7. Markov decision process is a reinforcement learning algorithm with applications in sequential decision-making context. MDP is one of the best tools for making sequential decision making under uncertainty.

## MULTIPLE CHOICE QUESTIONS

1. Stochastic process is a
  - (a) Collection of independent random variables
  - (b) Collection of random variables indexed by time  $t$
  - (c) Collection of random variables with stationary increments.
  - (d) Collection of random variables with independent increments
2. In a Poisson process the expected value  $\lambda t = 10$ . Then  $P[N(t) = 10]$  is
  - (a) 0.5
  - (b) 0.125
  - (c) 0.25
  - (d) 0.583
3. Arrival of patients at a hospital follows a Poisson process at a rate of 10 arrivals per hour. The time between arrivals is
  - (a) Normal distribution with a mean time between arrival of 6 minutes
  - (b) Uniform distribution with a mean time between arrival of 6 minutes

- (c) Poisson distribution with a mean time between arrival of 6 minutes  
 (d) Exponential distribution with a mean time between arrival of 6 minutes
4. For a positive recurrent state in a Markov chain  
 (a) Mean recurrence time is finite  
 (b) Mean recurrence time is infinite  
 (c)  $\sum_n P_{ii}^n > 1$ , where  $P_{ii}^n$  is  $n$ -step transition probability  
 (d)  $\sum_n P_{ii}^n < 1$ , where  $P_{ii}^n$  is  $n$ -step transition probability
5. For a transient state  $i$ , which of the following statements are true?  
 (a)  $f_{ii}^n < 1$ , where  $f_{ii}^n$  is the first passage time after exactly  $n$  steps      (b)  $f_{ii}^n = 1$   
 (c)  $F_{ii} < 1$  where  $F_{ii} = \sum f_{ii}^n$       (d)  $F_{ii} = 1$
6. A periodic Markov chain  
 (a) Has no limiting probability  
 (b) Has no stationary distribution  
 (c) May have limiting probability but no stationary distribution  
 (d) May have stationary distribution but no limiting probability
7. Stationary distribution of a Markov chain exists when  
 (a) The Markov chain is irreducible      (b) The Markov chain is aperiodic  
 (c) The Markov chain is irreducible and periodic      (d) The Markov chain is irreducible and aperiodic
8. Markov decision process is a  
 (a) Supervised learning algorithm      (b) Reinforcement learning algorithm  
 (c) Unsupervised algorithm      (d) Evolutionary algorithm

### EXERCISES

1. A sample data of 20 failures (time between failures) of laptop batteries (measured in months) is shown in Table 16.19.

TABLE 16.19 Time between failures

9	11	8	12	6	4	14	11	6	10
16	8	6	7	4	6	7	13	8	16

Assume that the data follows a Poisson process.

- (a) If the manufacturer provides 12 months warranty, what is the expected number of failures during the warranty period?  
 (b) The manufacturer would like to ensure they can meet the demand for spare batteries 90% of the times from the stock. How many batteries they should stock to meet this condition for 2-year period?
2. Demand for a bathroom fitting at a retail store follows a Poisson process with rate 3 per month. The cost of space occupied by the bathroom fitting per unit per month is INR 1200. The selling price of the bathroom fitting is 12000 and the margin (excluding the cost of space) is 4200. The retailer is planning to give a 10% discount on the unit price (selling price). What should be the increase in the demand rate so that the discount price is effective?
3. Customer's arrive at Bannerghatta petrol pump at 20 per hour and the amount of petrol they buy follows a normal distribution with mean 14.5 litres and standard deviation 3 litres. At the beginning of a day, the petrol pump has 5000 litres and the replenishment of petrol will arrive after 20 hours. Find the probability that the pump will go dry in 20 hours. Assume Poisson process for customer arrival at the petrol pump.

4. A fast-food restaurant provides unlimited soft drinks along with its meals. The time between arrivals of customers to the restaurant follows an exponential distribution and the mean time between arrivals of customers is 3 minutes. The quantity of drink consumed by customers has a mean of 300 ml and standard deviation of 120 ml. Calculate the probability that the demand for soft drinks over 10 hours will exceed 100 litres.
5. Customers are classified into the following three states (three different customer segments) by an e-commerce portal.

**STATE 1**


---

Low frequency Customer

---

**STATE 1**


---

Medium frequency Customer

---

**STATE 1**


---

High frequency Customer

---

At the starting of January 2017, the number of customers in three states is given in Table 16.20.

**TABLE 16.20** Distribution of customers

State	1	2	3
Customers at the beginning of January 1	500	250	250

The following transitions (Table 16.21) were observed from three states between January 2017 and February 2017.

**TABLE 16.21** Transition between states

	1	2	3
1	300	50	150
2	100	100	50
3	50	50	150

Construct a one-step transition probability matrix (TPM). Check whether the transitions follow a Markov chain using appropriate statistical test at 5% significance.

6. A firm classifies the customers into four customer segments: State 0 is a non-customer and states 1, 2, and 3 are low, medium, and high volume customers, respectively. The state transition between various states (constructed based on a quarterly data) is shown in Table 16.22

**TABLE 16.22** State transition diagram

	0	1	2	3
0	0.85	0.10	0.05	0
1	0.11	0.75	0.10	0.04
2	0	0.02	0.88	0.10
3	0.01	0.09	0.10	0.80

- (a) If a customer is currently in state 0, calculate the expected duration it will take the customer to reach state 3.
- (b) If a customer is in state 0 in the current period, what is the probability that he/she will be in state 3 after 2 periods?
- (c) Calculate the steady-state retention probability of the customers.
7. Five websites (labelled 1, 2, 3, 4, and 5) are being ranked based on the number links between them. Table 16.23 provides the details about existence of links from one website to another website.

**TABLE 16.23** Binary values indicating existence of a link between websites

	1	2	3	4	5
1	0	0	1	0	1
2	1	0	0	1	0
3	0	1	0	0	1
4	1	0	1	0	0
5	1	1	1	1	0

where,

$$a_{ij} = \begin{cases} 1, & \text{there is a link from website } i \text{ to website } j \\ 0, & \text{otherwise} \end{cases}$$

The transition probability between states is calculated using the following formula:

$$p_{ij} = \frac{a_{ij}}{\sum_{j=1}^5 a_{ij}}, \quad i = 1, 2, \dots, 5; j = 1, 2, \dots, 5$$

Construct a Markov transition probability matrix based on data in Table 16.21 and rank the websites using stationary distribution of Markov chain.

8. Number of weeks a movie is screened in multiplexes can be modelled using an absorbing state Markov chain, in which state 0 is the absorbing state (movie is removed from multiplexes) and state  $i$  ( $i = 1, 2, \text{ and } 3$  denotes  $i^{\text{th}}$  week in the multiplex). State 4 denotes greater than 3 weeks in multiplexes. The transition probability matrix is shown in Table 16.24.

**TABLE 16.24** Transition probability matrix

	1	2	3	4	0
1	0	0.20	0	0	0.80
2	0	0	0.60	0	0.40
3	0	0	0	0.70	0.30
4	0	0	0	0.20	0.80
0	0	0	0	0	1

- (a) Calculate the average life of movies in multiplexes using Markov chain model.
- (b) The average weekly collection is 0.15 million rupees. Calculate the expected future revenue for a movie that is in the second week at the multiplexes.
9. Bengaluru Consulting Group (BCG) classifies potential customers into different states as stated below.

**STATE 1**


---

Customer (that is, buys the consulting services)

---

**STATE 2**


---

Non-customer (informs that they are not interested in the consulting services)

---

**STATE 3**


---

New customer with no history

---

**STATE 4**


---

During most recent contact, customer's interest level was low

---

**STATE 5**


---

During most recent contact, customer's interest level was high

---

Based on the past data, the following transition matrix was derived (Table 16.25).

TABLE 16.25 State transition between different customer states					
	1	2	3	4	5
1	0.8	0.2	0	0	0
2	0	0.8	0	0.2	0
3	0	0	0	0.7	0.3
4	0	0.3	0	0.4	0.3
5	0.5	0.1	0	0.1	0.3

(a) For a new customer, determine the average number of contacts made before the customer buys the consulting service.

(b) What fraction of the new customers will buy the consulting services eventually?

10. Average weekly television rating points (TRP) of television programs are classified into 4 states as described below:

**STATE 1** TRP < 1**STATE 2**  $1 \leq \text{TRP} < 2$ **STATE 3**  $2 \leq \text{TRP} < 3$ **STATE 4**  $\text{TRP} \geq 3$

The average revenue generated in four states per minute of advertisement (in lakhs of rupees) is given in Table 16.26.

**TABLE 16.26** TRP and revenue

State	1	2	3	4
Revenue in Lakhs	2	4	6	8

In each state, the channel may decide to either promote or not promote the program. The cost of promotion in each state is dependent on the state due to the intensity of the promotion and is given in Table 16.27.

**TABLE 16.27** Cost of promotion in different TRP states

State	1	2	3	4
Cost of Promotion (in Lakhs)	2	2	1	1

The state transition between states under promotion is given in Table 16.28.

**TABLE 16.28** Transition matrix under promotion

	1	2	3	4
1	0.3	0.4	0.2	0.1
2	0.2	0.3	0.3	0.2
3	0.1	0.3	0.4	0.2
4	0.05	0.25	0.4	0.3

The state transition between states under no-promotion is given in Table 16.29.

**TABLE 16.29** Transition matrix under no promotion

	1	2	3	4
1	0.5	0.4	0.1	0
2	0.4	0.4	0.2	0
3	0.2	0.4	0.3	0.1
4	0	0.3	0.5	0.2

- (a) Using policy iteration algorithm of Markov Decision Process, check the better policy between the following two policies: 1. (P, P, P, NP) and 2. (P, P, NP, NP), where P implies promotion and NP implies no-promotion. Use a discount rate of 0.95.
  - (b) Find the optimal policy using linear programming formulation.
  - (c) Find the optimal policy for a finite horizon of 4 weeks.
11. An apparel retailer has to decide between different discount values during the end of season sale (EOSS) for its most popular brand to increase the sell-through rate. Sell-through rate is the percentage of item sold

(sell-through rate = items sold/items received). The retailer has defined the following three states (Table 16.30) for the sell through rate.

**TABLE 16.30** States based on sell-through rates

State	1	2	3
State Definition	Sell-through less than 50%	Sell-through between 50% and 75%	Sell-through more than 75%

The transition probability matrix measured in weeks (for various discount values) between different sell-through rate states are given in the following matrices (Tables 16.31–16.33).

**TABLE 16.31** Transition matrix with no discount (0%)

No Discount	1	2	3
1	0.75	0.2	0.05
2	0.4	0.5	0.1
3	0.2	0.3	0.50

**TABLE 16.32** Transition matrix with 20% discount

20% Discount	1	2	3
1	0.40	0.30	0.30
2	0.30	0.50	0.20
3	0.10	0.20	0.70

**TABLE 16.33** Transition matrix with 30% discount

30% Discount	1	2	3
1	0.20	0.40	0.40
2	0.20	0.50	0.30
3	0.10	0.10	0.80

The weekly rewards (in millions of rupees) under different states and discounts are given in Table 16.34.

**TABLE 16.34** Reward (measured in millions of rupees) for different states and discount

Discount	Reward under Different States		
	1	2	3
0%	7	9	12
20%	5	8	10
30%	4	6	8

- (a) Use policy iteration algorithm to choose the better policy between policies (0%, 20%, 30%) and (0%, 30%). Use a discount factor of 0.95.
- (b) Formulate a linear programming model to find the optimal policy.
- (c) In the current EOSS, only 3 weeks are left. Find the optimal policy that will maximize the reward for next 3 weeks of EOSS.

Case Study

## Customer Analytics at Flipkart.com<sup>1</sup>

It was typical cloudy monsoon weather at Bangalore on July 28, 2015. In the Darwin room of Flipkart's Cessna Business Park office, Ravi Vijayaraghavan, the Head of Analytics and Pravin Shinde, Senior Manager Analytics were brainstorming various business problems that Flipkart as an e-commerce company was encountering. Flipkart had been putting in much effort and emphasis on the use of analytics in every aspect of decision making. Forecasting demand for thousands of stock-keeping units (SKUs), predicting returns and cancellations of orders, predicting the reasons when customers contact the customer service centers, optimizing markdown pricing, identifying various types of frauds, optimizing vehicle routing, and enabling adherence to service-level agreements, were some of the typical problems that the analytics division of Flipkart was solving using state-of-the-art analytics techniques. In 2015, the team included about 100 data scientists mostly recruited from institutes such as the Indian Institute of Technology and Indian Institute of Management specifically for this purpose.

E-commerce in India had seen a compound annual growth rate (CAGR) of 34% since 2009 and was expected to exceed USD 22 billion by 2015 (**Exhibit 1**).<sup>2</sup> Under the e-commerce head, e-travel in itself comprised 71% of the total e-commerce market; e-tailing, which comprised online retail, and online marketplaces have been growing exponentially and are well-poised to become the fastest growing segment, expected to reach USD 56 billion by 2023 (**Exhibit 2**).<sup>3</sup> The industry believed that growth was at an inflection point with the key drivers being broadband internet, rising standards of living, wider product range, and changing lifestyles of Indian consumers. Such high growth rate also created several business challenges to e-commerce companies as well as to their marketplace suppliers, among them profitability still remained a major challenge. The e-commerce companies in India incurred combined losses of around INR 10 billion<sup>4</sup> through heavy discounting to penetrate into the brick-and-mortar retail customer base.

Ravi Vijayaraghavan started the meeting by stating:

We have been analysing our data to gain insights, but do we know the value of our customers? I think it is important for us to differentiate our customers through metrics such as customer lifetime value, which will help us to manage them effectively. For example, we can make our promotions effective if we know the customers with high customer lifetime value.

Customer lifetime value (CLV) is the net present value (NPV) of future cash flows (or profit). CLV is usually calculated at a customer segment level. The main challenge in calculating the lifetime value of cus-

<sup>1</sup> Copyright © the Indian Institute of Management Bangalore. The case is co-authored by Naveen Bhansali, Jitendra Rudravaram and Shailaja Grover and U Dinesh Kumar and is distributed through Harvard Business Publishing. This case is not intended to serve as an endorsement, source of primary data, or to show effective or inefficient handling of decision or business processes. Reproduced with the permission of IIM Bangalore.

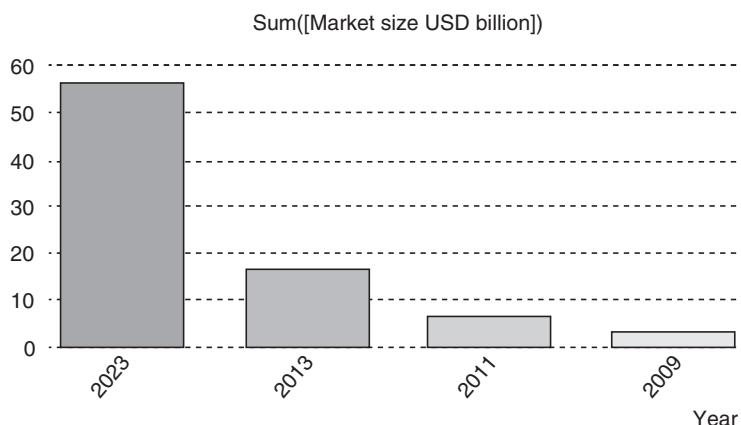
<sup>2</sup> Source: IndianOnlineseller.com - <http://indianonlineseller.com/2014/08/delhi-biggest-online-shoppers-1-4-indianbuys-mobile/>

<sup>3</sup> Source: Study of M&A scenario in Indian e-commerce market- <http://www.novonous.com/case-studies/analysis-ma-scenario-indian-e-commerce-market>

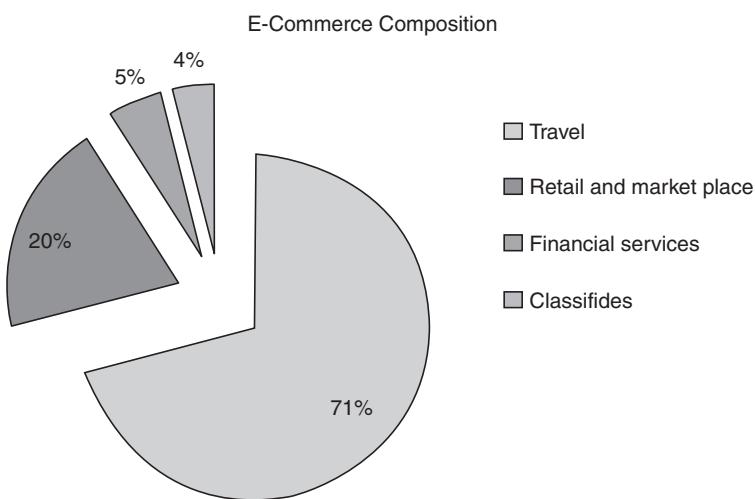
<sup>4</sup> Source: Live mint: <http://www.livemint.com/Industry/5hz6UnSAB9gaAeZ4OobwMM/Online-retailers-losses-total-Rs1000-crore-so-far.html>

Continued...

Case Study



**EXHIBIT 1** Growth of Indian e-commerce. Source: IndianOnlineseller.com - <http://indianonlineseller.com/2014/08/delhi-biggest-online-shoppers-1-4-indian-buys-mobile>.



**EXHIBIT 2** Composition of Indian e-commerce sector. Source: Study of M&A scenario in Indian e-commerce market- <http://www.novonous.com/case-studies/analysis-ma-scenario-indian-e-commerce-market>.

tomers of e-commerce companies such as Flipkart is that the exact life of the customer is unknown owing to data truncation; that is, the actual point in time of customer churn may not be identified in e-commerce, since there would be no prior communication from the customer about the churn. Hence, traditional models of CLV calculation may not be appropriate for e-commerce companies such as Flipkart.

## About Flipkart

Flipkart, the poster child of Indian e-commerce, was an early entrant in the nascent Indian e-commerce market and quickly established itself as the leading company in this space.

## Case Study Continued...

It was founded in 2007 by Sachin Bansal and Binny Bansal, both alumni of the Indian Institute of Technology, Delhi. They pooled in INR 2,00,000 (approximately USD 3,150) each to start Flipkart in 2007. From a startup with an investment of just INR 4,00,000 (approximately USD 6,300), Flipkart had grown into an online retail giant, valued at over USD 15.2 billion as of 2015. Flipkart was running the marathon with ample support from private equity players such as Tiger Global, which invested over USD 1 billion as of 2015.<sup>5</sup> Flipkart sold over 30 million products from more than 50,000 sellers in 70+ categories and consisted of 30 exclusive brand associations, with in-a-day guarantee in 50 cities and same-day guarantee in 13 cities. Flipkart was 33,000 people strong and had over 50 million registered users with over 10 million daily visits and 8 million shipments per month. A burgeoning consumer class, coupled with a rising web-literate population and zealous venture capital funding propelled Flipkart to become India's answer to Alibaba and Amazon.

The use of e-commerce to buy products and services was spreading at a fairly rapid pace in the psyche of the Indian consumer. In Indian cities such as Bangalore, lack of time, ease of shopping, and attractive pricing were major drivers for online shopping. On the other hand, accessibility to a variety of products encouraged customers from smaller towns and cities to opt for the online route.

### **Customer Analytics at Flipkart**

E-commerce companies such as Flipkart had access to huge amount of data, available for applying predictive and prescriptive analytics to take data-driven decisions. Flipkart had a strong analytics team headed by Ravi Vijayaraghavan, which used statistical models and machine learning algorithms to generate crucial customer insights. Flipkart had more than 50 million registered users and the transaction data of these customers could be used in a far more meaningful way using analytics to predict online consumer behaviour.

In 2015, the Indian e-commerce market space was facing immense competition owing to the entry of Jabong, HomeStop18, Infibeam, Indiaplaza, Snapdeal, and a plethora of other pure-play and multi-channel e-commerce companies. The continued growth of e-commerce and tough competition compelled Flipkart to seek a competitive advantage through more sophisticated analytics. Flipkart has been taking several analytics-enabled decisions, for instance, using web analytics to determine which landing pages encourage customers to make a purchase as well as which pay per click ad campaigns were most effective. In the face of tremendous competition, more than ever before, the analytics team at Flipkart aimed to predict customer demand for the products, understand its customer's loyalty, assess the true impact of customer retention strategies (discounts, coupons, and extra services), and focus on customer segments with higher retention and spend potential.

In 2015, Flipkart wanted to understand its customers better and retain most of them through effective promotions, since customer retention is less expensive as compared to customer acquisition.

<sup>5</sup> Source: [http://www.business-standard.com/article/specials/tiger-global-flipkart-s-largest-investor-and-second-largest-stockholder-in-amazon-115111701126\\_1.html](http://www.business-standard.com/article/specials/tiger-global-flipkart-s-largest-investor-and-second-largest-stockholder-in-amazon-115111701126_1.html)

**Continued...**

Unlike the churn in the telecom sector, which was clearly defined and captured (in the instance of postpaid customers), churn for e-commerce companies was difficult to define and capture, as these events were unobserved. Across e-commerce companies, the customer churn may be very high owing to reasons such as need fulfilment, cessation of demand, competition, and so on. However, it was important to capture customer churn and identify which customers should be retained.

### Churn Analysis and Lifetime Value

E-commerce companies faced a scenario of inconsistent customer purchase pattern wherein the gap between purchases could stretch far more than 6 months. Even though most of these buyers could come back after a gap of 5–6 months, Flipkart aimed to identify high-value customers, and subsequently increase purchase traction among them.

The analytics team at Flipkart wanted to model customer purchase patterns, repeat buyer trends, and calculate churn probabilities to help them identify the repeat customer segment to focus more on these customers for their marketing and promotional strategies. The final objectives of this exercise were to forecast the revenue generated from existing customers and calculate their lifetime value.

### Data Description

In order to carry out a detailed customer value assessment addressing customer churn issues, Flipkart collected sample transactional data spanning across 2 years: January 2013 to December 2014 such that all the 30,000 customers in the sample had made at least one purchase in January 2013. This was done to ensure new customers in the above said period were excluded from the study. Variable description of the data is provided in **Exhibit 3**.

**EXHIBIT 3** Unique Identifier: account\_id (Account id of the customer)

Variable	Description
item_selling_price	Selling price at the time of purchase
order_creation_date	Date when order was placed
Pincode	Address pin code
product_id	Product id of the product being purchased
unit_quantity	Quantity ordered
unit_status	Order delivery status

Source: Primary Data from Flipkart.

### Data Analysis

To understand customer churn and lifetime value, Pravin's team decided to use Discrete Time Markov Chains (DTMC). To build the churn model, the team first had to identify the period of inactivity (gap

**Continued...**

between transactions) to define churn. Gap was thus defined as the difference in months between two successive purchases or the difference between the current month (despite no purchase) and the last purchase month. The team put together the frequency distribution of all the purchases at different gaps, provided in **Exhibit 4**. This frequency table was used to define churn, that is, the absorbing state and built a DTMC as shown in **Exhibit 5**. The states of this Markov chain were defined on the basis of recency of purchase as shown in **Exhibit 6**. The Transition Probability Matrix (TPM) for this DTMC was computed using the transaction data and is shown in **Exhibit 7**.

Pravin's team also wanted to forecast revenue from existing customers and therefore built a model using Recency (defined as when was the last purchase was made by the customer) and Monetary (defined as how much money was spent in the latest month which had a purchase). The team retained the recency state definitions and augmented the state space by adding monetary slabs for each recency level. These monetary slabs were identified from the frequency distribution provided in **Exhibit 8**. The extended state space (combination of recency and frequency) definition is shown in **Exhibit 9**. Whenever a customer is in an inactive state, the monetary values are retained from the month of the last purchase. The TPM for the Recency–Monetary DTMC is provided as an accompanying file to the case. **Exhibit 10** includes the average monetary value for each of the earlier identified monetary slabs.

The team was also keen to identify customer segments based on their current transactional attributes such as recency, monetary, and frequency. The objective was to provide a fairly accurate mechanism to study the purchase patterns of various customer segments and thereby enable effective promotions to increase customer spend and arrest customer churn. The team chose a quarterly transition time period to account for macro-level stochastic changes in the model. **Exhibit 11** shows the frequency distribution for recency, frequency (defined as the number of distinct days in the quarter when a purchase was made), and monetary at a quarterly level. The results from **Exhibit 11** were used to define a RFM-based DTMC state space, shown in **Exhibit 12**. Whenever a customer is in an inactive state, the frequency and monetary values are retained from the quarter of the last purchase. **Exhibit 13** shows the customer sub-segments in the active state, that is, Recency 1. **Exhibit 14** shows the truncated TPM with states from 1 to 8 (active states) and states from 9 to 33 (inactive states) clubbed together.

The transition probability matrices from each of the models were used to calculate the customer lifetime value for customers in each segment and subsequently build an effective campaign strategy to reduce churn and increase customer spend.

**EXHIBIT 4** Gaps between transactions (measured in months) in purchase by customers

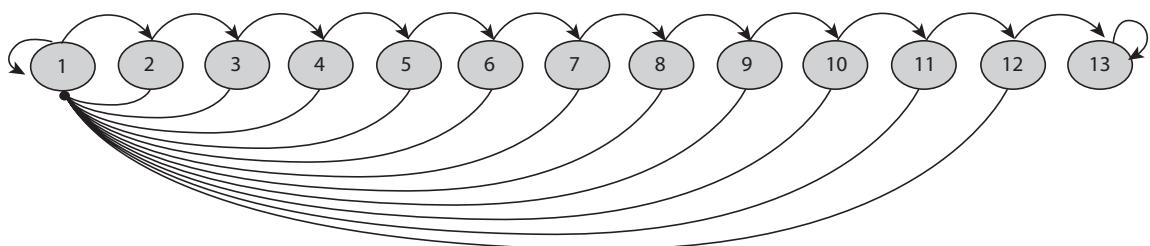
Gaps	Count of Purchases	%	Cumulative %
0	91806	52.80	52.80
1	31969	18.38	71.18
2	16423	9.44	80.63

(Continued)

**Continued...****EXHIBIT 4** Gaps between transactions (measured in months) in purchase by customers—Continued

Gaps	Count of Purchases	%	Cumulative %
3	9649	5.55	86.17
4	6263	3.60	89.78
5	4449	2.56	92.33
6	3165	1.82	94.15
7	2323	1.34	95.49
8	1638	0.94	96.43
9	1319	0.76	97.19
10	973	0.56	97.75
11	797	0.46	98.21
12	585	0.34	98.55
13	492	0.28	98.83
14	394	0.23	99.06
15	360	0.21	99.26
16	283	0.16	99.42
17	253	0.15	99.57
18	198	0.11	99.68
19	148	0.09	99.77
20	152	0.09	99.86
21	129	0.07	99.93
22	120	0.07	100.00
Total	173888		

Source: Based on the data provided by Flipkart.

**EXHIBIT 5** Transition diagram between recency states. Source: Based on the data provided by Flipkart.

**Continued...****EXHIBIT 6** Recency states

State	Recency Level	Explanation
1	1	Purchase made this month
2	2	Purchase made one month ago
3	3	Purchase made two months ago
4	4	Purchase made three months ago
5	5	Purchase made four months ago
6	6	Purchase made five months ago
7	7	Purchase made six months ago
8	8	Purchase made seven months ago
9	9	Purchase made eight months ago
10	10	Purchase made nine months ago
11	11	Purchase made ten months ago
12	12	Purchase made eleven months ago
13	13	Purchase made twelve months ago, hence churned

Source: Based on the data provided by Flipkart.

**EXHIBIT 7** One-step transition probability matrix (recency states)

States	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0.511	0.489	0	0	0	0	0	0	0	0	0	0	0
2	0.365	0	0.635	0	0	0	0	0	0	0	0	0	0
3	0.300	0	0	0.700	0	0	0	0	0	0	0	0	0
4	0.244	0	0	0	0.756	0	0	0	0	0	0	0	0
5	0.205	0	0	0	0	0.795	0	0	0	0	0	0	0
6	0.180	0	0	0	0	0	0.820	0	0	0	0	0	0
7	0.153	0	0	0	0	0	0	0.847	0	0	0	0	0
8	0.137	0	0	0	0	0	0	0	0.863	0	0	0	0
9	0.105	0	0	0	0	0	0	0	0	0.895	0	0	0
10	0.103	0	0	0	0	0	0	0	0	0	0.897	0	0
11	0.091	0	0	0	0	0	0	0	0	0	0	0.909	0
12	0.079	0	0	0	0	0	0	0	0	0	0	0	0.921
13	0	0	0	0	0	0	0	0	0	0	0	0	1

Source: Based on the data provided by Flipkart.

**Continued...****EXHIBIT 8** Monetary value definition

Monetary Amount	Frequency	Percentage	Cumulative Percentage
499	74787	36.75	36.75
999	45780	22.50	59.25
1999	36828	18.10	77.35
4999	28090	13.81	91.16
9999	9917	4.87	96.03
229093	8075	3.97	100.00
Total	203477	100	-

**EXHIBIT 9** State definition using recency and monetary

State	Recency Level	Monetary Level	Explanation
1	1	>9999	Purchase made this month for value higher than Rs. 9999
2	1	4999–9999	Purchase made this month for value between Rs. 4999–9999
3	1	1999–4999	Purchase made this month for value between Rs. 1999–4999
4	1	999–1999	Purchase made this month for value between Rs. 999–1999
5	1	499–999	Purchase made this month for value between Rs. 499–999
6	1	<499	Purchase made this month for value lower than Rs. 499
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
67	12	>9999	Purchase made twelve months ago, for value higher than Rs. 9999
68	12	4999–9999	Purchase made twelve months ago, for value between Rs. 4999–9999
69	12	1999–4999	Purchase made twelve months ago, for value between Rs. 1999–4999
70	12	999–1999	Purchase made twelve months ago, for value between Rs. 999–1999
71	12	499–999	Purchase made twelve months ago, for value between Rs. 499–999
72	12	<499	Purchase made twelve months ago, for value lower than Rs. 499
73	13		Purchase more than 12 months back hence churned

Source: Based on the data provided by Flipkart.

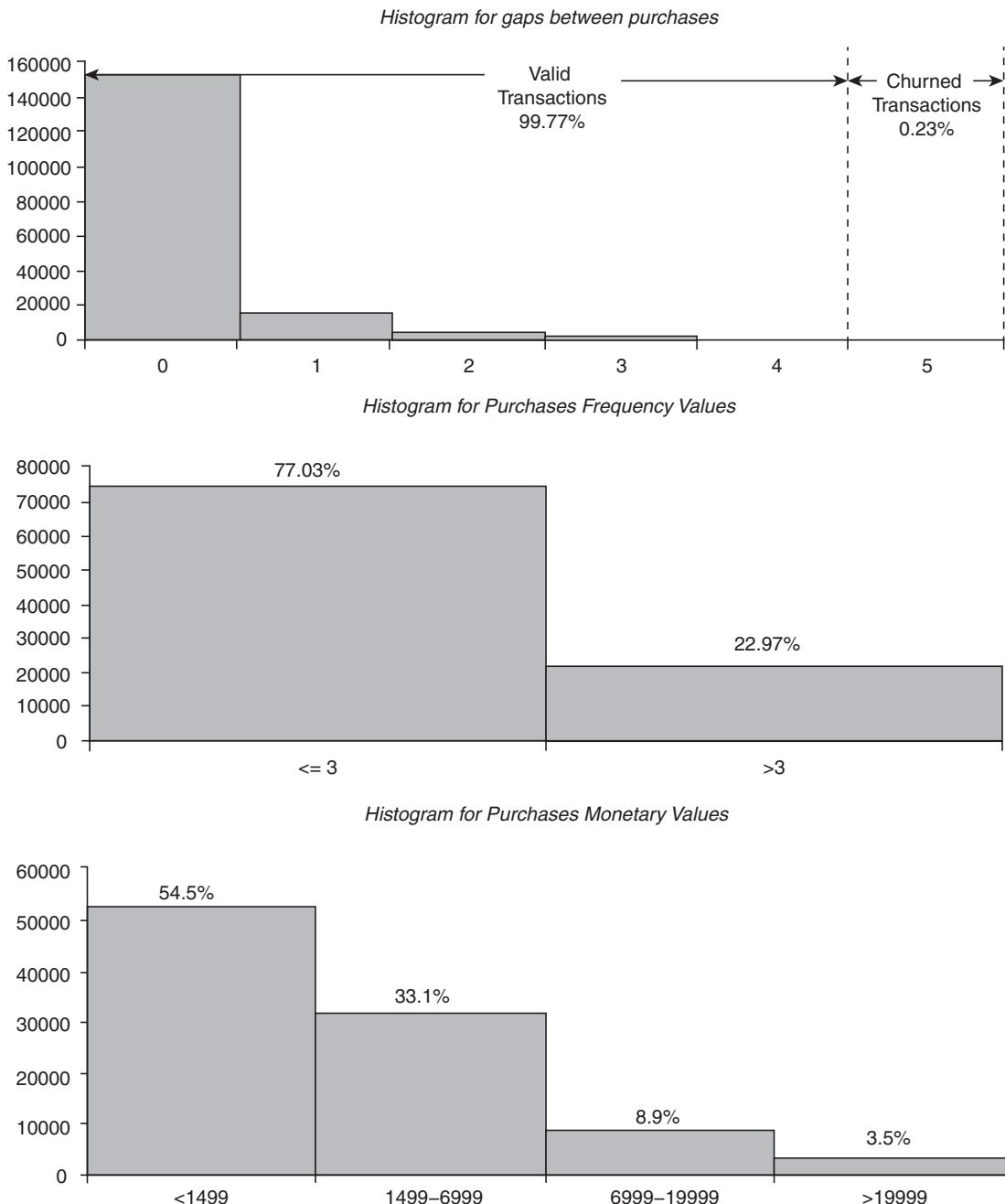
**EXHIBIT 10** Average revenue in monetary states

State	State 1	State 2	State 3	State 4	State 5	State 6
Average Revenue / Level	22032	6977	3114	1423	720	304

Source: Based on the data provided by Flipkart.

Case Study

**Continued...**



**EXHIBIT 11** Histogram of gaps (Measured in quarters) between purchases. Source: Based on the data provided by Flipkart.

Continued...

**EXHIBIT 12** RFM based states

State	R	F	M
1	1–3	>3	>19999
2	1–3	>3	6999–19999
3	1–3	>3	1499–6999
4	1–3	>3	<1499
5	1–3	≤3	>19999
6	1–3	≤3	6999–19999
7	1–3	≤3	1499–6999
8	1–3	≤3	<1499
9	4–6	>3	>19999
10	4–6	>3	6999–19999
11	4–6	>3	1499–6999
12	4–6	>3	<1499
13	4–6	≤3	>19999
14	4–6	≤3	6999–19999
15	4–6	≤3	1499–6999
16	4–6	≤3	<1499
17	7–9	>3	>19999
18	7–9	>3	6999–19999
19	7–9	>3	1499–6999
20	7–9	>3	<1499
21	7–9	≤3	>19999
22	7–9	≤3	6999–19999
23	7–9	≤3	1499–6999
24	7–9	≤3	<1499
25	10–12	>3	>19999
26	10–12	>3	6999–19999
27	10–12	>3	1499–6999
28	10–12	>3	<1499
29	10–12	≤3	>19999
30	10–12	≤3	6999–19999
31	10–12	≤3	1499–6999
32	10–12	≤3	<1499
33	13–15	All	All

Source: Based on the data provided by Flipkart.

Continued...

**EXHIBIT 13** Customer segments based on recency and frequency

Segments	Crème de la crème	High Spenders	Medium Spenders	Budget Buyers
States	15	26	37	48
Frequency	>3 ≤3	>3 ≤3	>3 ≤3	>3 ≤3
Monetary Spend Value	>19999	6999–19999	1499–6999	<1499
No of Customers*	458	1338	4439	6378
Average Revenue	39480 32273	11347 11326	3636 2984	1116 631

\*As per Q1 in the sample data set. Source: Based on the data provided by Flipkart.

**EXHIBIT 14** Transition probability matrix with RM states

States	Crème de la crème	High Spenders	Medium Spenders	Budget Buyers	Inactive
1	0.25	0.03	0.23	0.04	0.14
2	0.09	0.01	0.24	0.03	0.21
3	0.02	0.01	0.1	0.02	0.28
4	0.01	0	0.04	0.01	0.2
5	0.05	0.05	0.03	0.06	0.07
6	0.03	0.02	0.05	0.07	0.22
7	0.01	0.01	0.04	0.04	0.21
8	0.01	0.01	0.02	0.02	0.13
9	0.03	0.01	0.03	0.14	0.01
10	0.05	0.01	0.04	0.05	0.06
11	0.01	0.01	0.03	0.02	0.09
12	0	0	0	0.01	0.07
13	0.02	0.03	0.01	0.08	0.02
14	0.01	0.02	0.01	0.06	0.12
15	0.01	0.01	0.01	0.03	0.15
16	0	0.01	0.01	0.01	0.02
17	0	0.11	0.03	0.03	0.07
18	0	0	0.01	0.04	0.03
19	0.01	0	0.01	0.01	0.02
20	0	0	0	0.02	0.08
21	0	0.03	0	0.04	0.1
22	0	0.01	0	0.04	0.03
23	0	0.01	0	0.02	0.01

(Continued)

**Continued...****EXHIBIT 14** Transition probability matrix with RM states —Continued

	Crème de la crème	High Spenders	Medium Spenders	Budget Buyers	Inactive
24	0	0	0.01	0.01	0.05
25	0	0	0	0	0
26	0	0.01	0.01	0.05	0.01
27	0.01	0	0.01	0	0.06
28	0	0	0	0	0.07
29	0	0.04	0	0.06	0
30	0	0.01	0.01	0.03	0
31	0	0	0.01	0.01	0.01
32	0	0	0	0.01	0.04
33	0	0	0	0	0
					1

Source: Based on the data provided by Flipkart.

**CASE QUESTIONS**

1. Discuss whether the churn problem can be modelled as a Markov chain. What are the assumptions made while modelling customer churn as a Markov chain?
2. How can churn be defined for e-commerce companies such as Flipkart? In **Exhibit 6**, the recency state 13 is identified as the churn state. Comment on the use of state 13 as churn state.
3. Using **Exhibit 7** (Markov chain with recency states), Ravi and his team wanted to find out on average how many months customers in each non-absorbing state (states 1 to 12) take to reach the churn state (state 13).
4. Given a hypothetical case of 1,000 customers being in state 1, what would be the distribution of these 1,000 customers over a period of 4 months?
5. Given a hypothetical case of 1,000 customers in state 1, 1,000 customers in state 2, and 1,000 customers in state 3, predict the distribution of the customers after a period of 4 months from now? (Use **Exhibit 7**, recency state transition matrix.)
6. Ravi wants to evaluate Flipkart's relationship with a customer by calculating the expected life time value (CLV) for infinite horizon. Assuming Flipkart is risk neutral, and willing to make decisions based on expected net present value, calculate the CLV of a customer using **Exhibit 7** and information given below:

Discount rate  $d = 0.2$  and the reward in recency state (state 1) is 1000 and states 2 to 12 is -200  
(interpreted as cost of promotion) and state 13 is 0.

7. Karan and Arun are two Flipkart customers who made their first purchase in April 2013. Karan purchased products on Flipkart every month, except in August 2013, whereas Arun made his next purchase only in September 2013. From the months to churn (lifetime) calculated in Q2, calculate the estimated remaining lifetime for both Karan and Arun at the end of September 2013. (Round all decimals to the higher integer.)
8. The Analytics team built a Recency-Monetary Discrete Time Markov Chain model to predict revenue from existing customers using the information in **Exhibit 8**. The state space of the DTMC using recency-monetary information is described in **Exhibit 9**. The Transition Probability Matrix (TPM) of the Recency-Monetary DTMC is provided in a separate spreadsheet. Using the distribution of existing customers as of December 2014 (provided in the spreadsheet) for a particular segment and information in **Exhibit 10**, predict the number of customers in January 2015 and calculate the estimated revenue from this customer segment.
9. Using **Exhibit 14**, identify states where an intervention could be undertaken to reduce churn.

**REFERENCES**

1. Anderson T W and Goodman L A (1957), "Statistical Inference about Markov Chain", *The Annals of Mathematical Statistics*, **28**(1), 89–110.
2. Brin S and Page L (1998), "The Anatomy of a large-scale hypertextual web search engine", *Computer Networks and ISDN Systems*, **30**, 107–117.
3. Bellman R (1957), "Dynamic Programming", Princeton University Press, Princeton.
4. Bayes B (2013), "First Links in the Markov Chain", *American Scientist*, **101**, 91–97.
5. Cinlar E (1975). "Introduction to Stochastic Processes", Dover Publications, New York.
6. Haggstrom O (2007), "Problem Solving is Often a Matter of Cooking up an Appropriate Markov Chain", Chalmers University Report, available at [http://math.uchicago.edu/~shmuel/Network-course-readings/Markov\\_chain\\_tricks.pdf](http://math.uchicago.edu/~shmuel/Network-course-readings/Markov_chain_tricks.pdf), accessed on 15 May 2017.
7. Howard R A (1971), "Dynamic Probabilistic Systems Volume II: Semi-Markov and Decision Processes", John Wiley and Sons, New York.
8. Howard R (2002), "Comment on Origin and Application of Markov Decision Processes", *Operations Research*, **50**(1), 100–102.
9. Ross S M (2010), "Introduction to Probability Models", 10<sup>th</sup> Edition, Academic Press, USA
10. Styan G P H and Smith H (1964), "Markov Chain Applied to Marketing", *Journal of Marketing Research*, **1**(1), 50–55.



# 17

# Six Sigma

“Change before you have to.”

— Jack Welch

## LEARNING OBJECTIVES

- LO 17-1** Understand the role of analytics in process improvement and problem solving.
- LO 17-2** Understand how Six Sigma methodology is used for process improvement and problem solving.
- LO 17-3** Learn fundamental concepts in Six Sigma measures such as DPMO, Yield and Sigma Score. Understand the link between Six Sigma and process capability.
- LO 17-4** Understand Six Sigma process improvement methodology DMAIC.
- LO 17-5** Learn various tools and techniques used in DMAIC methodology.

## SIX SIGMA

Many business problems involve improving underlying processes. Business process improvement plays an important role for effective management and Six Sigma methodology is frequently used for process improvement. DMAIC (Define, Measure, Analyse, Improve, and Control) and DMADV (Define, Measure, Analyse, Design, and Verify) methodologies are frequently used for problem solving by improving the underlying processes and to design new processes and products.

### IMPORTANT

*Six Sigma is one of the most popular and frequently used analytics methodologies across several industries*

## 17.1 | INTRODUCTION TO SIX SIGMA

Six Sigma is one of the popular methodologies used across industries for solving problems and improving processes. One of the primary objectives of Six Sigma is improving customer satisfaction by ensuring defect-free delivery of products and services. A frequently quoted example of Six Sigma methodology in India is Mumbai Dabbawalas. Every morning in Mumbai, about 5,000 Dabbawalas (lunch box delivery men; Figure 17.1) collect about 200,000 *dabbas* (lunch boxes) from houses in

various suburbs of Mumbai, carry them on the suburban trains, and deliver them to various offices, colleges, and schools in and around Mumbai so that its citizens can eat fresh and hygienic homemade food (Thomke and Sinha, 2013). In 2017, the customers are charged approximately INR 600–1000 (around USD 10 to 15) per month for this service. The most amazing fact about this massive logistics operation is that the dabbawalas almost never fail to deliver the lunch boxes to their rightful owners. The dabbawala service, which started in 1890, continues to attract the attention of the academic world as well as the industry. In 1998, Forbes magazine reported that the reliability of this service in the delivery of dabba to their rightful owners meets Six Sigma standards (Chakravarty, 1998; Moore 2011).



**FIGURE 17.1** Mumbai dabbawala.

The defect rate (failure to deliver a lunch box to its rightful owner) of the dabbawalas is approximately 1 in 16 million transactions (Chakravarty, 1998). This puts them at an enviable position in the Six Sigma scale of performance – the Sigma Score of the dabbawalas, assuming one defect in 16 million deliveries, is 5.286 to be precise; even aircraft manufacturers such as Airbus and Boeing who have very high quality standards for every part used in an aircraft will be happy with such high Sigma Score. This high quality of service by dabbawalas is achieved using simple logistics processes. They use a simple colour and number coding system (Thomke, 2013) to track the dabba (lunch boxes). There are no computer-controlled systems – such as those used by logistics companies such as FedEx and UPS – to track the location of the lunch boxes during their 3-hour transit from Mumbai houses through the suburban railways of Mumbai to their owners. Many members of the 5000-strong dabbawala workforce are illiterate. The entire logistics of their operation is coordinated by the brain (or rather 5000 brains) of these dabbawalas using an effective colour and number coding system.

Many companies across various industries use Six Sigma to improve processes and solve customer problems. The primary objective of Six Sigma is not the achievement of a Sigma Score of 6 but rather continuous improvements that result in improved customer satisfaction and profitability.

Take the case of 3M, for instance: it started Six Sigma initiatives within its organization in February 2001, completed about 16,000 Six Sigma projects by August 2004, and had 16,000 on-going Six Sigma projects in 2004 (Anon, 2004). Companies such as Motorola and General Electric attribute their success to their Six Sigma programmes. For the sake of comparison, the delivery quality achieved by the Mumbai dabbawalas can be compared to the quality in delivering checked-in baggage of some leading U.S. airlines (Table 17.1, based on data cited in Bowen and Headley, 2002). It should be noted that the baggage handling system used by these airlines uses some of the most sophisticated technologies. However, as far as the quality of delivery is concerned, they have a lot to catch up on and their delivery quality is far behind that of the dabbawalas (Table 17.1). The good news is that the safety and reliability of air travel in general are much higher than Seven Sigma, making it one of the safest modes of transport. Thanks to Boeing and Airbus, passengers are flown to their destinations at a higher sigma score.

**TABLE 17.1** Sigma level of checked-in baggage handling of major U.S. Airlines

Airline	Number of Mishandled Baggage Items per 1000 checked-in Baggage Items	Sigma Level
American Airlines	4.60	4.10
Continental Airlines	4.29	4.13
Delta Air Lines	4.11	4.14
Northwest Airlines	4.19	4.14
Southwest Airlines	4.77	4.09
Trans World Airlines	6.35	3.99
United Airlines	5.07	4.07

Six Sigma is philosophically different from earlier quality and process improvement techniques. Prior to Six Sigma, the defects were defined as non-conformance to specification; however, in Six Sigma anything that results in customer dissatisfaction is treated as defect. Table 17.2 lists the number of complaints received about different airlines in India and the corresponding Sigma Score. The data from airlines across the world shown in Tables 17.1 and 17.2 exhibits the achievement of Mumbai Dabbawalas. Airlines use highly sophisticated systems to manage their business processes.

**TABLE 17.2** Number of complaints about different domestic airlines in India and corresponding Sigma Score (March 2013)

Airline	Number of Complaints per 10,000 Passengers	Sigma Score
Air India	1.8	3.57
Jet Airways	1.6	3.60
Jet Lite	1.4	3.63
IndiGo	1.3	3.65
SpiceJet	1.2	3.67
GoAir	0.4	3.94

Source: [http://dgca.nic.in/reports/pass\\_complaints.pdf](http://dgca.nic.in/reports/pass_complaints.pdf).

## 17.2 | WHAT IS SIX SIGMA?

Six Sigma is a structured problem-solving methodology aimed at improving customer satisfaction and company profitability by improving underlying processes. Dinesh Kumar *et al.* (2006) defined Six Sigma as a ‘management strategy which provides a roadmap to continuously improve business processes to eliminate defects in products, processes, and service’. The core of Six Sigma methodology is process improvement. A common misconception among academics and practitioners is that Six Sigma initiatives are meant only for solving quality issues in manufacturing. Although Six Sigma originated as a tool to reduce the failure rate in Motorola, it can be used in many other contexts. For example, it can be used for improving the mean time between failures (MTBF) of a system, efficiently managing crowd at airports and railway stations, traffic congestions, minimizing waste in a restaurant, reducing the waiting time in banks, reducing errors in financial reporting, improving on-time delivery in a supply chain systems, improving marketing effectiveness, improving student performance, and so on.

Charles Duhigg (2012), in his book *The Power of Habit*, narrates an incident in which a U.S. army major used an innovative method to reduce riots in Iraq. After analysing video footage of the riots, he identified a pattern in the formation of crowds and the initiation of riots. He ordered the removal of kebab vendors from these gatherings. This ensured that no food was available at such gatherings, which led to the dispersion of the crowds because people became hungry after some time and decided to go home. The major did not use the Six Sigma methodology explicitly to solve this problem. However, the approach he took has significant resemblance to the Six Sigma methodology. He collected data (video footage of public gatherings that later turned into riots), identified the vital cause (availability of food, which allowed people to stay on until someone threw a stone or some other object), and removed the vital cause (the kebab vendors). The important learning from this example is that Six Sigma is not meant merely for fixing quality issues. It is a structured problem-solving methodology with wider applications.

Six Sigma uses Sigma Score as one of the measures to measure the performance of a process and effectiveness of a process in solving customer problems. Each problem is treated as a project and either DMAIC (Define, Measure, Analyse, Improve, and Control) or DMADV (Define, Measure, Analyse, Design, and Verify) methodology is used for solving the problem. DMAIC is used to improve an existing process whereas DMADV is used to create a new process.

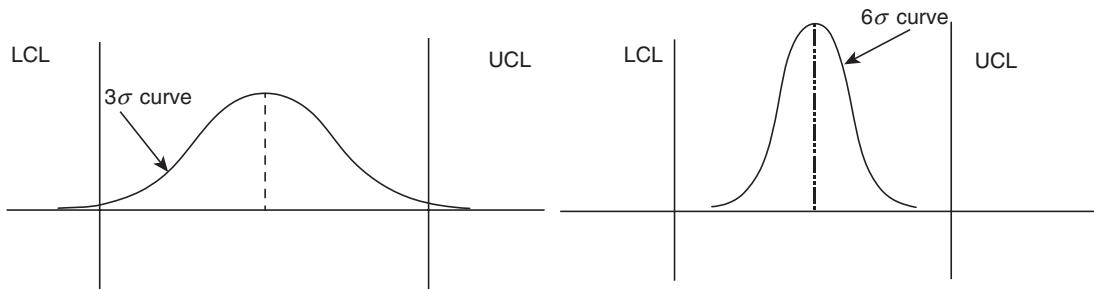
## 17.3 | ORIGINS OF SIX SIGMA

For many decades now, quality has been an important discriminating factor across industries. Several strategies have evolved over the last several decades to improve quality in the manufacturing and service industries. Six Sigma was introduced in the 1980s by William Smith of Motorola as a method to reduce manufacturing defects. Motorola was one of the first two companies that won the Malcolm Baldrige National Quality Award in 1988 for developing what later came to be known as the Six Sigma methodology (Anon, 1988). The Six Sigma methodology tries to achieve a process capability index value of 2; prior to the Six Sigma era, a process capability index value of 1 was an acceptable quality standard.

The term ‘Six Sigma’ was coined by William Smith, a reliability engineer at Motorola, and is a registered trademark of Motorola. Smith pioneered the concept of Six Sigma to deal with the higher than

expected failure rate experienced by the systems developed at Motorola. Smith proposed Six Sigma as a goal for improving the reliability and the quality of products. Until then, the lower specification limit (LSL) and the upper specification limit (UCL) for processes were fixed at three-sigma ( $3\sigma$ ) deviations from the mean. Smith suggested that these limits (LSL and USL) should be pushed to six-sigma ( $6\sigma$ ) levels. That is, the LSL and USL should be set at  $\mu - 6\sigma$  and  $\mu + 6\sigma$ , where  $\mu$  is the process mean and  $\sigma$  is the standard deviation of the process. This would force the designers to design their processes with minimum deviations (minimum process variation).

In simple terms, quality can be defined as *conformance to specification*. However, not many would accept such a simple definition – today, anything that results in customer dissatisfaction is considered as a defect and is treated as an indication of bad quality. As per the ‘conformance to specification’ definition of quality, specifications are usually defined using lower and upper specification limits. Prior to the Six Sigma methodology, the lower specification limit (LSL) and the upper specification limit (USL) for a process were usually set at three-sigma ( $3\sigma$ ) levels:  $\mu - 3\sigma$  (LSL) and  $\mu + 3\sigma$  (USL). Figure 17.2 presents a comparison of three-sigma and six-sigma processes for given mean, LSL, and USL. In Figure 17.2, the six-sigma curve is much narrower (better controlled due to less variability) compared to three-sigma curve.



**FIGURE 17.2** Comparison of three-sigma and six-sigma processes.

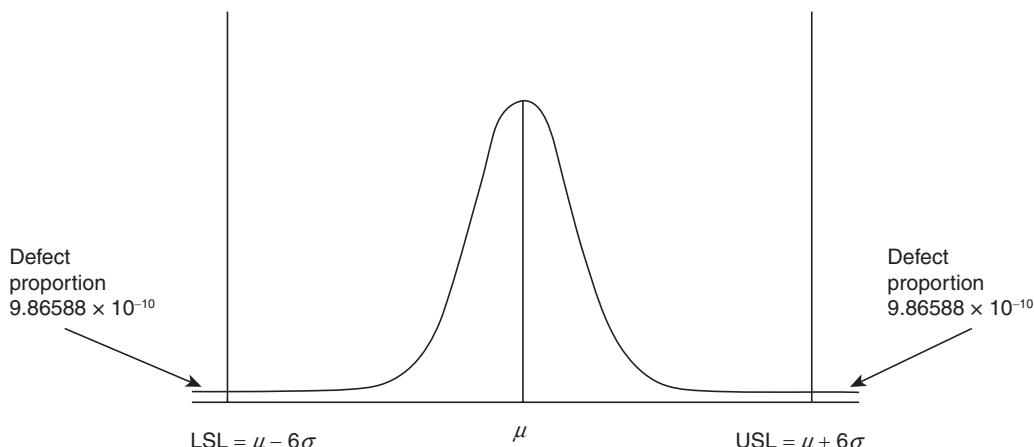
Reducing process variations is one of the core objectives of Six Sigma projects, since process variations result in greater loss to corporates. Taguchi and Clausing (1990) reported a classic example of the impact of process variations using the case of Ford versus Mazda. Ford, which owned a 25% stake in the Japanese company Mazda, had asked the latter to build the transmissions (gear-boxes) for the cars that Ford sold in the United States. The transmissions were built to specifications identical to those used by Ford; Ford had adopted zero defects as its standard. However, after the cars had been in the market for some time, it was observed that Ford’s transmission systems were generating far higher warranty costs as compared to the transmission systems built by Mazda. The reason was traced to the fact the Ford’s transmissions had higher process variability compared to the transmissions built by Mazda. Sony Corporation reported something similar for their televisions manufactured at Tokyo and San Diego. The televisions manufactured at Tokyo had less variability in colour density compared to the televisions manufactured at San Diego. As a result, the customer satisfaction level for the televisions manufactured at Tokyo was much higher than that for the televisions manufactured at San Diego.

Six Sigma can be used as a manufacturing strategy to reduce the number of defects as well as a business strategy to improve business processes and to evolve new business models. Many proponents of Six Sigma stress that the power of Six Sigma lies in the fact that it can be used as a business strategy for improving market share and profitability. The Design for Six Sigma (DFSS) concept uses Six Sigma as a strategy for designing and developing new products or for revamping an existing process; the traditional Six Sigma methodology aims to reduce defects. The methodology adopted by DFSS is called DMADV – Design, Measure, Analyse, Design, Verify.

## 17.4 | THREE-SIGMA VERSUS SIX-SIGMA PROCESS

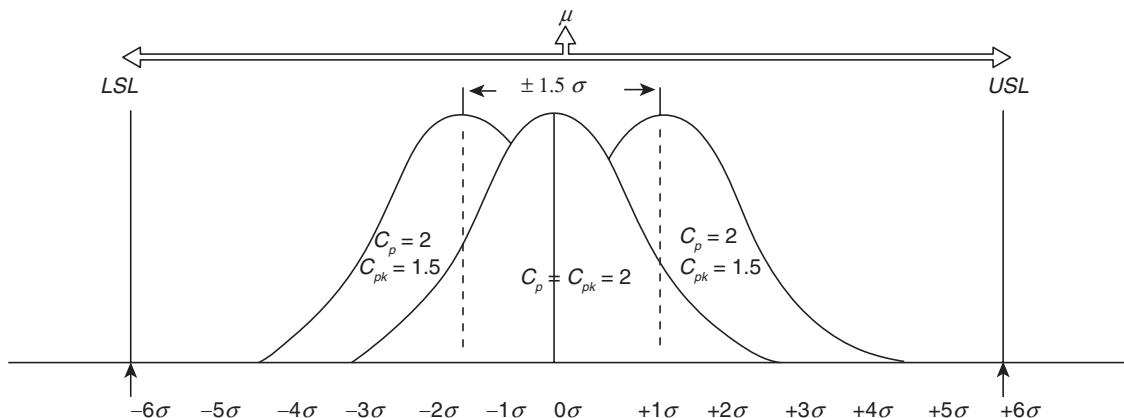
Lower and upper specification limits act as the threshold for process parameter. Consider a company that sells packed milk of 500 ml. It is almost impossible to pack exactly 500 ml in each packet. Let LSL be 490 ml and USL be 510 ml with a mean ( $\mu$ ) of 500 ml. If LSL and USL are set at three-sigma (that is, LSL at  $\mu - 3\sigma$  and USL at  $\mu + 3\sigma$ ) then the value of  $\sigma$  is  $10/3$ . In a six-sigma process, the lower and upper specification limits are set at  $\mu - 6\sigma$  and  $\mu + 6\sigma$ , respectively. The corresponding  $\sigma$  is  $10/6$ . That is, when the mean is at the centre, moving from a three-sigma process to a six-sigma process will involve reducing the process variation to half. Most importantly, changing a process from  $3\sigma$  to  $6\sigma$  has a significant impact on the number of defective parts per million (PPM) opportunities.

Once LSL and USL are set, any output that lies beyond LSL and USL is treated as defect. Six Sigma drives business processes to achieve a defect rate of less than 2 (1.9731 to be precise) Defects per Billion Opportunities (DPBO). Using the example of 500 ml milk packet described earlier in which  $\mu = 500$ ,  $LSL = 490$ , and  $USL = 510$ , setting the LSL and USL limits at six standard deviations ( $6\sigma$ ) from the mean of a normal distribution would yield a defect rate of approximately 2 DPBO (Figure 17.3). The area to the left of LSL in Figure 17.3 is  $9.86588 \times 10^{-10}$  under the normal distribution mean  $\mu = 500$  and standard deviation  $\sigma = 10/6$ . Similarly, the area to the right of USL is  $9.86588 \times 10^{-10}$ . Thus, the total defective proportion is approximately 2 DPBO.



**FIGURE 17.3** Defect proportion under Six Sigma specification limits.

One of the issues in process management is that the process mean itself may shift. For example, if the packaging system of the milk is not maintained well, the mean fill of 500 may shift to a different value due to deterioration in the packaging system. Key performance indicators (termed as *critical to quality* in Six Sigma methodology) may shift over a period of time due to natural deterioration. Researchers at Motorola found that the process mean may shift by as high as  $1.5\sigma$  (Figure 17.4). When the process mean shifts by  $1.5\sigma$ , the process defects will be 3.4 DPMO for a Six Sigma process. This assumption of process shift may be true in some cases, but is definitely not true for all processes.



**FIGURE 17.4**  $1.5\sigma$  shift in process mean.

The assumption that the process mean may drift by  $1.5\sigma$  over time has received considerable attention and criticism in the literature. One reason ascribed for this shift in the process mean is the learning effect or entropy. Consider an automobile company that manufactures cars, for instance. Fuel efficiency of the car is one of the important parameters of interest to customers. An improvement in the fuel efficiency of the car can be expected over a period of time due to the learning that happens within the company once the data about the processes become available, leading to improvements in some of the processes. Suppose a car model is released with a mileage of 20 km per litre, with standard deviation of 2 km. Over time, the mileage of the cars developed subsequent to this model is likely to increase due to improvements in the design and manufacturing processes. On the other hand, the mileage of an individual car is likely to decrease over a period of time due to wear and tear. A car model with a mileage of 20 kilometres per litre of fuel in the first year of manufacturing may achieve lower mileage, say 15 kilometres per litre after 10 years due to deterioration of parts. Magnitude of shift in process mean need not be  $1.5\sigma$ , it can be different for different processes.

## 17.5 | COST OF POOR QUALITY

Cost of poor quality (CoPQ) measures the cost of the resources that are used for activities that exist as a result of process deficiencies. The CoPQ is the sum of conformance cost and non-conformance cost, where conformance cost is the cost related to the prevention of poor quality and non-conformance cost is the cost resulting from the poor quality of a product and/or service failure. Pyzdek (2015) claimed that typical companies that operate between three- and four-sigma levels spend 25–40% of their revenues

fixing quality-related problems. Most of these costs are referred to as hidden costs, which are buried in the general operating costs. Hidden costs would include rework, customer complaints, warranty repairs, quality rechecks, and training. A company operating at a six-sigma level spends about 1–2% on quality Pyzdek (2015). This is primarily spent on quality resources – Black Belts and Green Belts, customer surveys, and metrics evaluation.

Any improvement in the sigma level of a company is likely to reduce the CoPQ. The CoPQ as a result of manufacturing defects is a function of rework cost, excessive use of material, warranty related costs, and excessive use of resources. Assuming that the CoPQ is linearly related to the number of defects produced by a process, the benefit of the relative increase in sigma level decreases (Kumar *et al.*, 2008). Improving a process beyond a certain sigma score (say beyond 4.5 sigma) may involve replacing the existing process with a new process or technology.

## 17.6 | SIGMA SCORE

The quality of a process can be measured using Sigma Score, also known as Sigma quality level. Table 17.3 presents the number of failures/defects per million opportunities for various sigma levels. The DPMO values in Table 17.3 were derived assuming that the process variation is normally distributed and there is no shift in the process mean. Table 17.4 shows the number of defects per million opportunities under the assumption that the process mean itself can shift up to  $1.5\sigma$ .

**TABLE 17.3** Sigma level and defects per million opportunities (DPMO) without shift in process mean

Sigma Score	DPMO
1	317,300
2	45,500
3	2700
4	63
5	0.57
6	0.002

**TABLE 17.4** Sigma level and defects per million opportunities (DPMO) with  $1.5\sigma$  shift in process mean

Sigma Score	DPMO
1	697,672
2	308,770
3	66,811
4	6,210
5	233
6	3.4

Figure 17.5 shows the relationship between sigma score and decrease in the number of defects as a function of process yield. In economic terms, Figure 17.5 illustrates that because of the decreasing returns to scale in process yield as sigma quality increases, at some point, it may not be economically beneficial to increase the Sigma Score, especially if the required process change calls for high investment. For example, consider improving a process from a three-sigma level to a four-sigma level. This improvement would reduce the number of DPMO from 2700 to 63. From a Taguchi quality loss<sup>1</sup> perspective (Taguchi, 1986), at the three-sigma level, the Taguchi quality loss would be higher compared to the Taguchi quality loss at the four-sigma level. Similarly, improving the sigma level from a five-sigma level to a six-sigma level would reduce the DPMO from 0.57 to 0.002. However, in most cases, more effort would be required to improve a process from a five-sigma to a six-sigma level as compared to the effort required for improving a process from a three-sigma to a four-sigma level. The improvement of the sigma score would necessarily require a reduction in the process variability, that is, the process sigma itself.

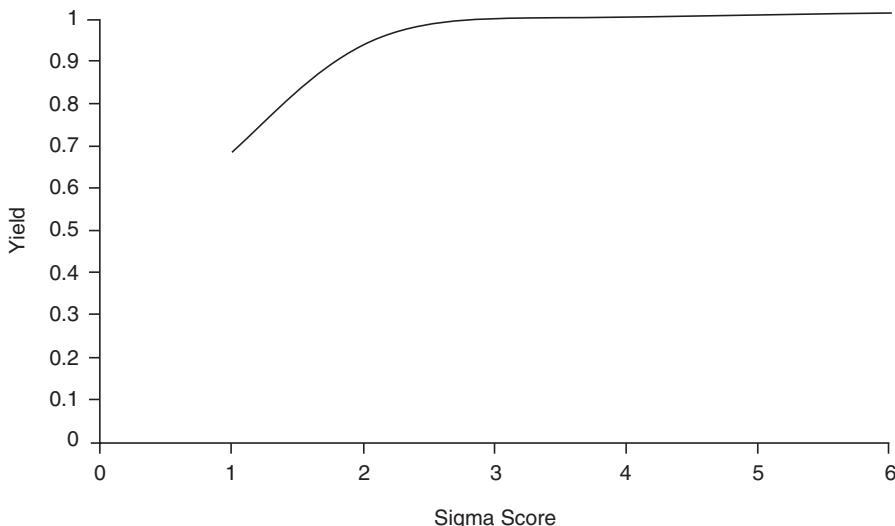


FIGURE 17.5 Decreasing returns to scale in process yield as sigma score increases.

Sigma Score can be used for benchmarking purposes and helps to measure the quality of a process. It also helps to set a realistic target for the improvement of process quality during the DMAIC cycle of process improvement.

## 17.7 | INDUSTRIAL APPLICATIONS OF SIX SIGMA

Six Sigma was initially designed for the manufacturing sector and for a while, it was perceived to be relevant only for this sector. Over time, as Six Sigma evolved, it began to be used across different sectors.

<sup>1</sup> The loss incurred due to deviation from target,  $L$ , is modelled as a quadratic function and is given by  $L = K(y - t)^2$ , where  $K$  is a constant,  $y$  is the achieved performance value, and  $t$  is the target.

The popularity of Six Sigma is mainly due to its applications across different industries. Several companies from various industries such as healthcare, finance, software development, food processing, business process outsourcing (BPO), automobile, insurance, semiconductor, and telecommunications regularly use Six Sigma to improve their processes. Today, it is hard to find any industry where Six Sigma is not used. Wyper and Harrison (2000) presented a case study involving Six Sigma deployment in human resource management. The scope of Six Sigma projects ranges from minor improvements such as reducing the failure rate of components to much larger scope such as increasing customer satisfaction and profitability. The list of companies that benefitted from using the Six Sigma methodology is quite large. Table 17.5 presents a sample list of the industrial applications of Six Sigma as reported in various studies.

**TABLE 17.5** Industrial applications of Six Sigma

Reference	Industrial Application
Hendricks and Kelbaugh (1998)	Successful implementation of several Six Sigma projects at GE that improved net profit
Lanyon (2003)	Improvement of HR process using Six Sigma
Motwani et al. (2004)	The Dow Chemical Co., which implemented Six Sigma on a corporate-wide basis in 2000, achieved its target of USD 1.5 billion in cumulative earnings before interest and taxes
Knowles et al. (2004)	Successful application of Six Sigma within the UK-based confectionery plant of a major food producer
Edgeman et al. (2005)	Claims savings of USD 2–3 million at the Office of the Chief Technology Officer (OCTO), Washington DC using Six Sigma strategies
Liu (2006)	Presented an application of Six Sigma to reduce the cycle time and the defects in clinical report entry
Mukhopadhyay and Ray (2006)	Used Six Sigma to reduce defects while packing yarn

The success of any methodology is measured by its contribution to the company's bottom line. Motorola reported savings of USD 840 million in three years using Six Sigma (Hahn et al., 1999). GE reported savings of over USD 300 million in its 1997 annual report. Interestingly, GE's stock price doubled within five years of introducing Six Sigma in the company (Lee and Choi, 2006). Samsung reported savings of USD 40 million within six months of the introduction of Six Sigma. Table 17.6 lists a few of the benefits due to Six Sigma within the manufacturing and service sectors that have been reported in the literature.

**TABLE 17.6** Benefits of Six Sigma in manufacturing sector

Company/Project	Metric/Measures	Benefits/Savings
Motorola (1992)	In-process defect levels	150 times reduction
Raytheon: Aircraft integration system	Depot maintenance inspection time (measured in days)	Reduced 88%
GE: Railcar leasing business	Turnaround time at repair shops	62% reduction
AlliedSignal (Honeywell): Laminates plant in South Carolina	Capacity; cycle time; inventory; on-time delivery	Up 50%; down 50%; down 50%; increased to nearly 100%
AlliedSignal (Honeywell): Bendix IQ brake pads	Concept-to-shipment cycle time	Reduced from 18 months to 8 months

**TABLE 17.6** Benefits of Six Sigma in manufacturing sector—Continued

Company/Project	Metric/Measures	Benefits/Savings
Hughes Aircraft's Missile Systems Group: Wave soldering operations	Quality; productivity	Improved 1000%; improved 500%
Continental Teves: Brake and axle assemblies	Failure rate	More than 50% reduction in failure rate
BorgWarner Turbo Systems	Financial	USD 1.5 million annually since 2002
General Electric	Financial	USD 2 billion in 1999
Motorola (1999)	Financial	USD 15 billion over 11 years
Dow: Rail delivery project	Financial	Savings of \$2.45 million in capital expenditures
DuPont: Yerkes plant in New York (2000)	Financial	Savings of more than USD 25 million
Telefonica de espana (2001)	Financial	Savings and increase in revenue: Euro 30 million in first 10 months
Texas Instruments	Financial	USD 600 million
Johnson and Johnson	Financial	USD 500 million
Honeywell	Financial	USD 1.2 billion
Ford Motor Company: Exterior surface defects	Financial	USD 500,000

Sources: Kwak and Anbari, (2006); Weiner (2004); De Feo and Bar-El (2002), Antony and Banuelas (2002); Buss and Ivey (2001); McClusky (2000); U Dinesh Kumar *et al.* (2008).

Six Sigma as a methodology has its own share of limitations. Potential users need to understand these limitations before implementing this methodology in their organization. According to the results of a survey involving major aerospace companies (conducted by *Aviation Week* magazine), less than 50% of the participating companies expressed satisfaction with the results of Six Sigma projects, nearly 20% were 'somewhat satisfied', and around 30% were dissatisfied (Zimmerman and Weiss, 2005). Even at these levels of satisfaction, Six Sigma was found to do better than many other process improvement techniques. Zimmerman and Weiss (2005) noted that 60% of the companies in the survey had selected opportunities for improvement on an ad-hoc basis and only 31% had relied on a portfolio approach. Interestingly, the companies that had used a portfolio approach had gained better results.

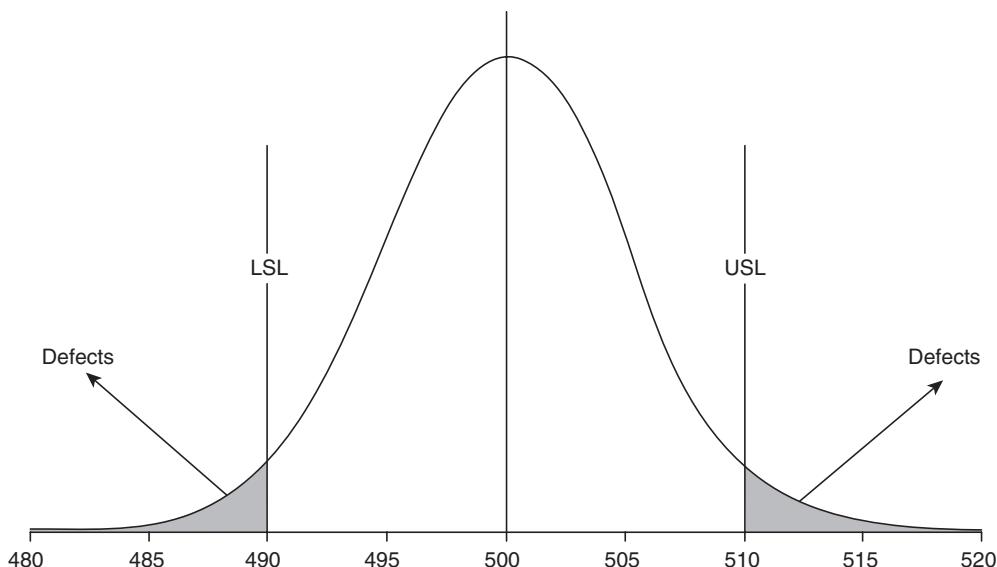
## 17.8 | SIX SIGMA MEASURES

### 17.8.1 | Process Capability and Process Capability Indices

Process capability is the ability of the process to conform to the specification and is a measure of process performance. The fundamental difference between previous quality initiatives and Six Sigma is the target set by Six Sigma for the process capability. For a Six Sigma process, the target process capability value is 2. Prior to Six Sigma a process capability index value of 1 was acceptable standard across several industries.

One of the basic steps in process design is the definition of lower specification limit (LSL) and upper specification limit (USL). LSL and USL are defined for key performance indicators (KPIs) of process,

product, and service. Six Sigma uses the term critical to quality (CTQ) for important performance indicators. LSL is the lowest acceptable value of the metric used to capture the process performance. If an output has a value less than LSL, then it will be classified as a defect. Consider the example of packaging milk packets discussed earlier. It is very difficult to achieve a process that can pack exactly 500 ml of milk all the time. Now the question is what should be the minimum quantity of milk that the process should achieve. Let this minimum value be 490 ml, that is at most 10 ml less than the target. One can argue why 490 ml, why can't we fix the minimum value as 480 ml? Well, the decision to arrive at a minimum value for the process itself would require extensive research. In this particular case, one can argue that most customers would not notice if the actual content is 490 ml instead of 500 ml and even if they notice they are unlikely to go to consumer courts. Any milk packet that contains less than 490 ml will be treated as a defect. Similarly, the maximum excess milk that can be packed in any packet is say 10 ml, that is, the USL in this case would be 510 ml. The company will incur loss if the actual content is more than 500 ml, since the customers are charged only for 500 ml. Defect in this example is any packet with milk content either less than 490 ml (LSL) or more than 510 ml (USL). If we assume that the actual quantity of milk in any packet is a random variable and follows normal distribution, then Figure 17.6 shows the distribution of quantity of milk packed by this process. In Figure 17.6, it is assumed that the process mean is 500 ml and standard deviation ( $\sigma$ ) is 5. The important issue to be noticed here is that the LSL and USL are not fixed arbitrarily and the values are driven by customer, market, or design. For example, instead of milk, suppose if someone is buying 500 grams (gm) of gold. The customer would not accept even one-millionth of a gram less than 500 gm. Similarly, the seller is unlikely to provide more than 500 grams of gold.



**FIGURE 17.6** Process capability chart for a milk packaging process with a target of 500 ml, LSL = 490 and USL = 510.

Process capability,  $C_p$ , is the ability of the process to produce products within the set specification limits. The mathematical expression for process capability index,  $C_p$ , is given by

$$C_p = \frac{USL - LSL}{6\sigma} \quad (17.1)$$

Equation (17.1) assumes that the process target is same as the mean ( $\mu$ ) value (500 grams) and is at the centre of LSL (490 grams) and USL (510 grams) values. For the aforementioned example, the process capability,  $C_p$ , is given by ( $\sigma = 5$ )

$$C_p = \frac{USL - LSL}{6\sigma} = \frac{510 - 490}{6 \times 5} = \frac{20}{30} = 0.667$$

Since  $\mu = 500$  and  $\sigma = 5$ , LSL = 490 is at  $\mu - 2\sigma$  and USL = 510 is at  $\mu + 2\sigma$ .

If we assume that LSL =  $\mu - 6\sigma$  and USL =  $\mu + 6\sigma$ , then the value of  $C_p$  is

$$C_p = \frac{(\mu + 6\sigma) - (\mu - 6\sigma)}{6\sigma} = \frac{12\sigma}{6\sigma} = 2$$

Thus, for a Six Sigma process the value of process capability index,  $C_p = 2$ . For many processes, the mean value may not be at the centre of the spread between LSL and USL. In this case, we define the process capability index  $C_{pk}$ , which is given by

$$C_{pk} = \text{Min}\left(\frac{USL - \mu}{3\sigma}, \frac{\mu - LSL}{3\sigma}\right) \quad (17.2)$$

For example, let us assume that LSL = 490, USL = 510,  $\mu = 505$ , and  $\sigma = 5$ . In this case, the value of  $C_{pk}$  is given by

$$C_{pk} = \text{Min}\left(\frac{USL - \mu}{3\sigma}, \frac{\mu - LSL}{3\sigma}\right) = \text{Min}\left(\frac{510 - 505}{15}, \frac{505 - 490}{15}\right) = 0.33$$

In many cases, only one specification limit will be valid. For example, consider service organizations such as bank, where waiting time to complete the service is one of the critical to quality parameters. Usually when parameters such as waiting times are involved, one would specify only USL. In such cases, only the part involving USL in Eq. (17.2) will be used. Equations (17.1) and (17.2) assume that the mean of the process is the target, which need not be true. If the target ( $T$ ) is different from mean ( $\mu$ ), then the process capability index  $C_{pm}$  is given by

$$C_{pm} = \frac{USL - LSL}{6\sqrt{\sigma^2 + (\mu - T)^2}} \quad (17.3)$$

Equation (17.3) integrates Taguchi quality loss function concept while calculating process capability.

$C_p$ ,  $C_{pk}$  and  $C_{pm}$  represent process capability indices and their usage depends on conditions that exist in a particular process. Table 17.7 summarizes the underlying assumptions of different forms of process capability indices and the relationship with other indices.

**TABLE 17.7** Process capability indices

Index	Underlying Assumptions	Relationship with Other Indices
$C_p$	The process mean ( $\mu$ ) is centered Process mean is same as the target ( $T$ ) value of the process	$C_p \geq C_{pk}$ $C_p \geq C_{pm}$
$C_{pk}$	The process mean is not centered Process mean is same as the target ( $T$ ) value of the process	$C_{pk} \leq C_p$
$C_{pm}$	The process mean is centered Process target is different from the process mean	$C_{pm} \leq C_p$

### 17.8.2 | Difference between Specification Limits and Control Limits

Specification limits are the limits of the variation that do not result in a defect. Specification limits describe whether the deviation in the process is acceptable to the internal and external customers. In this case, internal customers are the other departments within the company. Specification limits are best explained in a manufacturing scenario. Referring back to the earlier example of milk packets, if the quantity of the milk is less than 490 ml, the customer would be able to tell the difference in the quantity and if the quantity is more than 510 ml, the manufacturer would incur a loss. Hence in this case the USL and LSL are 510 ml and 490 ml, respectively, which are defined by the customer and dairy. This is the range which is acceptable for both the external customer and internal customer.

Process control limits or simply control limits are generated from the actual performance of the process. It is important for the user to apply statistical concepts to determine the control limits to ensure that the process is under control. One of the simplest ways of calculating the Upper Control Limit (UCL) and Lower Control Limit (LCL) is by calculating the mean and standard deviation of the process. Assume that the estimated mean from a sample is 500 and the estimated standard deviation is 2. Then the three-sigma UCL and LCL are  $506 (\mu + 3\sigma)$  and  $494 (\mu - 3\sigma)$ , respectively. That is, the specification limits are specified by the customers and/or design, whereas control limits are derived from the data to check whether a process is in control or out of control using statistical process control techniques such as control charts.

### 17.8.3 | Importance of USL and LSL

One frequent argument in Six Sigma literature is that it is possible to make any process into a Six Sigma process. For example, consider a hypothetical courier company that promises to deliver the documents within a specified duration. Currently, the average time taken to deliver the document is 18 hours and the standard deviation is 2 hours. If the company defines the defect as 'any document that takes more than 36 hours to deliver', then USL is within the Six Sigma limit ( $\mu + 6\sigma = 18 + 6 \times 2 = 30$ ), and the process will be a Six Sigma process. However, if the defect is defined as 'any document that takes more than 24 hours to deliver', then the current process is not a Six Sigma process. If the tolerance limits (LSL and USL) are large enough (further away from process target), then the corresponding process Sigma score will be high. However the problem is that organizations cannot

fix the tolerance levels arbitrarily. Tolerance levels are driven by the design, market, and customers. If a courier company promises to deliver the documents within 24 hours and do a good job of it, then there will be an expectation from customers that any courier company should be able to deliver within 24 hours. The following example establishes the importance of tolerance levels, and how stringent it has to be in many cases.

Indian Space Research Organization (ISRO) is one of the successful organizations in the space research across the world today. However, like any high technology organization, it had its share of failures and challenges and one such failure was the failure of GSLV (Geosynchronous Satellite Launch Vehicle) on 10<sup>th</sup> July 2006. ISRO's GSLV-F02 satellite launch vehicle crashed into Bay of Bengal shortly after its launch from Sriharikota on 10<sup>th</sup> July 2006. GSLV-F02 was carrying the communication satellite INSAT-4C. The cost of launch vehicle is estimated at 150 crore (approximately \$37 million) and the cost of satellite is estimated at 100 crore (approximately \$25 million). During the investigation the main cause of the failure was traced to regulators in one of the engines. The regulator in one of the engines had a bore of 17 mm instead of 16 mm (Anon, 2006). The regulators were purchased from an Indian vendor at the cost of INR 100,000. The cost of this 1 mm difference was approximately \$62 million. In some situations, the design tolerances are very stringent that even Six Sigma tolerance levels may not be sufficient; one may have to set the USL and LSL at higher sigma deviation. The important point to note here is that the tolerance levels such as USL and LSL are driven by many factors and the process designer cannot set USL and LSL arbitrarily; for the same reason, it is not correct to argue that any process can be converted into a Six Sigma process by adjusting LSL and USL.

### EXAMPLE 17.1

Vastra is an apparel manufacturer and one of their products is XL size shirt for which the chest size is 42 inches. The LSL and USL for XL size shirt are set at 41.5 and 42.5 inches, respectively, by the company. A sample of XL size shirts manufactured by Vastra had a mean chest size of 42 inches with standard deviation of 0.1 inch. Calculate the process capability index for the XL shirts manufactured by Vastra.

**Solution:** Since the mean value is same as the process target, the process capability index  $C_p$  is given by

$$C_p = \frac{USL - LSL}{6\sigma} = \frac{42.5 - 41.5}{6 \times 0.1} = \frac{1}{0.6} = 1.667$$

### EXAMPLE 17.2

Thickness of a computer chip used in a big data server should be 17.75 mm and the corresponding LSL and USL are 17.10 mm and 18.50 mm respectively. A sample of 50 chips was tested for its thickness and the data is shown in Table 17.8. Calculate the process capability for this process.

**TABLE 17.8** Thickness of 50 computer chips

16.55	17.82	17.55	18.59	19.97
17.13	16.48	17.02	18.48	17.12
17.74	17.52	18.51	17.60	16.21
17.61	19.70	17.24	17.71	17.98
16.00	16.45	17.92	16.35	16.07
19.47	16.85	18.03	18.98	19.18
16.11	16.68	18.80	19.11	18.26
16.78	18.64	18.23	19.54	19.46
18.79	16.56	19.64	19.50	16.17
16.05	16.55	16.88	16.07	19.85

**Solution:** From the data we get that the average thickness of chips is 17.75 mm and the standard deviation is 1.2272. Since the average thickness is not at the center of LSL and USL, we use the following equation to calculate the process capability index:

$$C_{pk} = \text{Min}\left(\frac{\text{USL} - \mu}{3\sigma}, \frac{\mu - \text{LSL}}{3\sigma}\right) = \text{Min}\left(\frac{18.5 - 17.75}{3 \times 1.2272}, \frac{17.75 - 17.10}{3 \times 1.2272}\right) \\ = \frac{17.75 - 17.10}{3 \times 1.2272} = 0.1765$$

## 17.9 | DEFECTS PER MILLION OPPORTUNITIES (DPMO)

Defects per million opportunities (DPMO) is an important measure in Six Sigma methodology and measures the number of defects in a process in terms of million opportunities for committing defect. While using Six Sigma, it is important to realize different complexity levels within the process. For example, manufacturing an aircraft is more complex than manufacturing a car. Thus, the number of opportunities for committing defect in manufacturing aircraft will be much higher than that of a car. Defect, on the other hand, is any process output that does not meet the customers' expectations. It is also defined as a non-conformance of a quality characteristic (e.g., waiting time, taste, ambience, etc. in case of a restaurant) to its specification. A product can have multiple defects and such a product will be deemed to be a 'defective product'. DPMO is measured in millions for convenience. It is also known as DPPM (defects counted in parts per million). The mathematical expression for DPMO is given by

$$\text{DPMO} = \frac{\text{Number of defects}}{\left( \frac{\text{Number of opportunities}}{\text{for defects per unit}} \right) \times (\text{Number of units})} \times 10^6 \quad (17.4)$$

The number of opportunities for defects per unit measures the possible number of defects per unit. For example, in the case of Bombay Dabbawalas, the tiffin carrier collected from a residence changes hands five times before it reaches the rightful owners. Thus, the number of opportunities for defects per unit in this case is 5. In many cases, the DPMO is measured using the ratio of number of defects over the number of opportunities. Usually the DPMO is estimated from a sample of units produced. DPMO is calculated from estimating the defects per unit (DPU) using the following expression:

$$DPU = \frac{\text{Number of defects}}{\left( \begin{array}{l} \text{Number of opportunities} \\ \text{for defects per unit} \end{array} \right) \times (\text{Number of units})} \quad (17.5)$$

Thus, DPMO is given by

$$DPMO = DPU \times 10^6 \quad (17.6)$$

One drawback of the DPMO definition in Eq. (17.6) is that it fails to differentiate the criticality of various defects. The Ford versus Mazda case discussed earlier highlights the importance of differentiating the defects based on criticality. For example, defects in few cases can be reworked to remove the defect, whereas in other cases it may not be possible to rework and the defective product may have to be scrapped. The Taguchi's quality loss function clearly differentiates the cost of poor quality based on the deviation from the target.

### EXAMPLE 17.3

Partha Diagnostics is a famous diagnostic centre in Bangalore. Their main service is to conduct various tests on blood samples of patients and produce reports based on the test. The diagnostic process and the opportunities for error for each of the activities in the process are shown in Table 17.9. Out of 4568 blood sample tests conducted in the previous year, 268 errors were found. Calculate the DPMO and DPU for this process.

**TABLE 17.9** Blood test process

Activity	Opportunities for Defects in the Activity
Collection of blood sample in a container from the patient and labeling the blood sample with patient name and the type of test	2
Transfer of blood sample from the container to test equipment and conducting the test identified in the blood sample container	3
Generating the report with patient details based on the output generated by the test equipment	2
Delivery of the report to the patient on time	1

**Solution:**

The number of opportunities for error per unit = 8

Number of units = 4568

$$\text{DPMO} = \frac{\text{Number of defects}}{\left( \begin{array}{l} \text{Number of opportunities} \\ \text{for defects per unit} \end{array} \right) \times (\text{Number of units})} \times 10^6$$

$$= \frac{268}{8 \times 4568} \times 10^6 = 7333.6252$$

$$\text{DPU} = \frac{\text{DPMO}}{10^6} = 0.0073$$

## 17.10 | YIELD

Yield ( $Y$ ) is an important measure of process capability and is defined as the ratio of the total number of units free of any defects to the total number of units produced, expressed in percentage. Mathematically, Yield can be written as

$$\text{Yield } (Y) = \frac{\text{Number of opportunities} - \text{Number of defective units}}{\text{Number of opportunities}} \times 100\% \quad (17.7)$$

For example, if 900 units were produced and there were 36 defective units, then Yield is given by

$$Y = \frac{900 - 36}{900} \times 100\% = \frac{864}{900} \times 100\% = 96.00\%$$

Yield can also be obtained using the following relationship:

$$\text{Yield} = (1 - \text{DPU}) \times 100\% \quad (17.8)$$

Since many defective units in manufacturing can be reworked to remove defects, organizations also calculate measures such as first time yield (yield without any rework) and final yield (yield after rework).

### EXAMPLE 17.4

Calculate the yield for Example 17.3.

**Solution:**

$$\text{Yield} = (1 - \text{DPU}) \times 100\% = (1 - 0.0073) \times 100\% = 99.27\%$$

It is possible that total number of defective units may be less than the number of defects; we may have to account for this while calculating yield which is a ratio of defective-free units to the total units.

### 17.10.1 | Rolled Throughput Yield

Rolled throughput yield is used when several processes are involved in manufacturing a product. For example, if  $n$  different processes are used to manufacture a product, then the rolled throughput yield ( $Y_{RT}$ ) is given by

$$Y_{RT} = \prod_{i=1}^n Y_i \quad (17.9)$$

where  $Y_i$  is the process yield of process  $i$ .

### 17.11 | SIGMA SCORE (OR SIGMA QUALITY LEVEL)

Sigma Score or Sigma quality level is a measure of process performance used by Six Sigma methodology. Very often students confuse between Sigma Score and Sigma (standard deviation). Sigma ( $\sigma$ , standard deviation) measures the variability in the process, higher sigma implies poorly managed process. Sigma Score measures the performance of the process which can be used to compare process performance across organizations. In Six Sigma, we try to decrease sigma (standard deviation) as low as possible, whereas we try to increase Sigma Score as high as possible.

Sigma score is basically Z-score under a standard normal distribution for which the area under the distribution is equal to the process yield.

Sigma Score is also a measure of process success rate. A higher sigma level indicates that the process results in fewer defects, whereas a lower sigma means higher defect rate. Sigma Score can be used for benchmarking purpose and helps to measure quality of the process. Sigma Score also helps to set a realistic target for improvement of the process quality during the DMAIC cycle of process improvement. Sigma Score can be calculated from DPMO as well as from Yield.

#### 17.11.1 | Conversion of Yield to Sigma Score under No Shift in the Process Mean

Let  $Y$  denote yield of a process represented in percentage. As discussed earlier, Sigma Score is nothing but the  $Z$ -value of the standard normal distribution for which the area under the standard normal distribution is equal to the yield (if there is no shift in the process mean). The Sigma Score  $Z$  is given by

$$Z = F^{-1}\left(\frac{Y}{100}\right) \quad (17.10)$$

where  $F^{-1}(Y/100)$  is the inverse cumulative standard normal distribution function for a given yield, also known as quantile function. In Microsoft Excel, Normsinv(Y/100) function will give the Sigma Score for a given yield when there is no process shift.

### 17.11.2 | Conversion of DPMO to Sigma Score under No Shift in Process Mean

If DPMO of a process is known then the corresponding Yield can be estimated using the following function:

$$Y = \left( 1 - \left( \frac{DPMO}{10^6} \right) \right) \times 100\% \quad (17.11)$$

The corresponding Six Sigma can be calculated using Eq. (17.10). In Eq. (17.11) we count the defective parts per million input units.

### 17.11.3 | Sigma Score under Process Shift

When there is a shift in the process mean by a  $K_p$  standard deviation, then the corresponding Sigma Score is given by (if yield,  $Y$ , is calculated in percentage)

$$Z = K_p + F^{-1} \left( \frac{Y}{100} \right) \quad (17.12)$$

The corresponding Microsoft Excel function is  $K_p + \text{Normsinv}(Y/100)$  or  $\text{Norminv}(Y/100, K_p, 1)$ .

#### EXAMPLE 17.5

For the Partha diagnostics problem, calculate the Sigma score without shift for the mean and with 1.5 sigma shift for mean.

**Solution:** The yield for Partha diagnostics is 99.27%. The corresponding sigma level without any process shift is given by the Excel function:

$$\text{NORMSINV}(0.9927) = 2.44$$

The sigma level with 1.5 sigma shift is given by

$$1.5 + \text{NORMSINV}(0.9927) = 3.94$$

#### EXAMPLE 17.6

In a hospital, all the food orders placed by inpatients are to be served within 20 minutes. Any order which is served after 20 minutes is treated as defective. The data collected over one month is analysed for the time taken to serve the food. The mean time taken is 14 minutes and the standard deviation is 2 minutes. Calculate the sigma level.

**Solution:** The Yield is  $\Phi((20 - 14)/2) = 0.99865$ .  $\Phi(\cdot)$  is the cumulative distribution function of standard normal distribution. The corresponding Sigma Score value is 3 ( $\text{NORMSINV}(0.99865)$ ).

## 17.12 | DMAIC METHODOLOGY

DMAIC framework is a step-by-step iterative procedure, successively used during problem solving by improving underlying processes. DMAIC is the acronym for Define, Measure, Analyse, Improve, and Control – a structured problem-solving methodology used for improvement of existing processes in which each problem or improvement opportunity is treated as a project. Working through the five phases DMAIC allows the managers to obtain solution to the problem at hand. Project management approach is used at each phase of the DMAIC. Project-specific goals are defined and achieved using appropriate tools. DMAIC is a proven way for solving business problems where the solution to the problem is unknown. The methodology enables continuous improvement that is determined by measurable critical business needs or key performance indicators. Depending on the scope of the projects, Six Sigma projects are classified as ***green belt projects*** and ***black belt projects***. Green belt projects are small improvement projects in which the project can be completed within three to six months using simple tools, whereas, black belt projects are big improvement projects requiring more skilled project team using sophisticated analytical tools to solve the problem. The main goals of each phase of DMAIC are listed in Table 17.10.

**TABLE 17.10** Goals of different stages of DMAIC

Define	Identify a problem where the solution is unknown. The problem is always defined in terms of critical to quality (CTQ or set of CTQs).
Measure	Collect data and establish a baseline Sigma Score or other macro/micro level Six Sigma matrices, the current state of CTQ, and extent of problem.
Analyze	Establish the relationship between the CTQ and critical to processes (CTPs). Identify the root causes of the problem.
Improve	Generate solutions either to remove root causes or reduce the impact of root causes, select the cost-effective solution and pilot it, measure the improvement.
Control	Develop the sustenance plan, sustain the gains, and convert the improvement into procedures

Table 17.11 shows activities and deliverables during various stages of Six Sigma.

**TABLE 17.11** List of activities and deliverables during various stages of Six Sigma

	<b>Define</b>	<b>Measure</b>	<b>Analyse</b>	<b>Improve</b>	<b>Control</b>
Activities	1. Identify project team. 2. Identify problem and/or improvement opportunity as a function of critical to quality (CTQ). 3. Identify problem statement, goals in terms of improvement in CTQs.	1. Clearly define defect and opportunities. 2. Collect data and measure process capability, DPMO, Yield and baseline Sigma Score. 3. Estimate any other project specific metric (such as waiting time).	1. Identify root causes of performance gap. 2. Establish the relationship between CTQ and CTP.	1. Generate solutions to eliminate causes identified in Analyze stage. 2. Perform cost effectiveness solutions. 3. Define and validate hypotheses on vital causes.	1. Identify success factors of project. 2. Develop and deploy a sustenance plan. 3. Validate ideas through proof of concept.

(Continued)

**TABLE 17.11** List of activities and deliverables during various stages of Six Sigma—Continued

	<b>Define</b>	<b>Measure</b>	<b>Analyse</b>	<b>Improve</b>	<b>Control</b>
	4. Develop high-level process map. 5. Develop project plan and milestones.	4. Develop detailed process map.			
Deliverables	1. Project plan (Green Belt or Black Belt Project)  2. Targets for CTQs  3. Project Team  4. Process Map	1. Relevant data.  2. Detailed process map.  3. Set realistic improvement targets.  4. Current process capability, DPMO, Yield and Sigma Score.	1. Vital causes for performance gap  2. Opportunities for improvements through CTPs.	1. Cost effective solutions.  2. Deployment plan.	1. Sustenance and monitoring strategy.  2. Project document.

Each stage of DMAIC methodology is discussed in detail in the following sub-sections.

### 17.12.1 | Define Stage

Define is the phase when a problem is identified or an improvement opportunity is identified and scoped. DMAIC starts with defining the problem in terms of CTQs. For example, CTQs could be waiting time at a bank, time taken to discharge a patient in a hospital, on-time delivery by courier and e-commerce companies, Net Promoters Score (NPS), mean time between failures (MTBF) and so on. Identification of CTQs are carried out using voice of customers (VOC) and high level process mapping technique such as Suppliers, Inputs, Process, Output, Customers (SIPOC).

### 17.12.2 | Measure

In measure stage we collect data to understand the current process performance using current Sigma score and value of other CTQ measures relevant for the problem identified in define stage (such as average waiting time and current MTBF). Critical measures that are necessary to evaluate the success of the project (such as financial impact of the DMAIC project if successful) are identified and determined. The metrics chosen should also facilitate measurement of critical to quality (CTQ) by identifying important input and output variables.

### 17.12.3 | Analyse

The primary objective of analyse stage is to identify the vital few causes that are responsible for decreased process performance or decreased Sigma Score. This phase is important in pinpointing and verifying

causes affecting the input and output variables related to the project goals. A detailed data analysis is carried out to reduce many reasons of a problem occurrence to the critical few causes. A relationship between the critical to process (CTP) to the critical to quality (CTQ) will be established. Apart from these, activities such as Pareto chart, cause-and-effect study, time and motion analysis, 5 Whys, statistical data analysis such as hypothesis testing and regression will be performed in order to aid the identification of the vital few root causes. The root causes are prioritized for actions during improve stage.

#### 17.12.4 | Improve

In this stage ideas are generated to eliminate (or reduce the impact of) vital causes responsible for lower Sigma Score or lower performance. There could be several potential solutions; the cost-effective solution is chosen for implementation.

#### 17.12.5 | Control

After the implementation of the solution we should make sure continuous improvement in the project/process is maintained. That is, a sustenance strategy has to be designed to ensure that the improvements obtained using solutions generated in improve stage are sustained. The key to successful control stage is designing an effective monitoring plan to sustain the performance improvements.

### 17.13 | SIX SIGMA PROJECT SELECTION FOR DMAIC IMPLEMENTATION

The success of DMAIC will depend on choosing the right problem/project. Although Six Sigma is a successful methodology, there are several instances where Six Sigma did not yield desired performance. Thus, it is necessary we understand how we would decide whether the identified problem is a candidate for the application of DMAIC. The most important characteristic is that the solution of the problem has significant impact on business. The impact may be to improve customer satisfaction, profit, preserve the business market position. The general guidelines are

1. Project should have identifiable process inputs and outputs.
2. If the solution is already known then just go for it instead of going through the DMAIC.
3. All projects need to be approached from with an objective of understanding the variation in process inputs, causes of variation controlling or managing variation, and eliminating the causes that result in defects.
4. Use techniques such as analytics hierarchy process to rank the projects based on multiple criteria for implementation of DMAIC.

### 17.14 | DMAIC METHODOLOGY – CASE OF ARMOURED VEHICLE

In this section, we will be discussing a case study of DMAIC deployment using engineer tank (armoured vehicle). The case is based on the material presented by Kumar UD *et al.* (2006). To maintain the confidentiality, the names of the user and the manufacturer are not disclosed. Engineer armoured vehicles, also known as Engineer Tank, are used on the battlefield to undertake engineering tasks. This vehicle's

primary task on the battlefield is to open routes and provide mobility to the army. Its tasks include providing a route across short gaps, countering mines by clearing routes across mine fields, obstacle breaching and clearance (by digging and bulldozing). The functional elements for the armoured vehicle are: mobility, surveillance, communications, lethality, and survivability. During the development of the engineer tank engine, reliability targets were set and these targets are shown in Table 17.12

**TABLE 17.12** Target reliability for the engineer tank engine

Factor	Engine Hours
Mean Time Between Failure (MTBF) [Failures which result in unscheduled maintenance]	420
Mean Time Between Mission Critical Failures (MTBMCF) [Failure which cause loss of mobility]	3500
Mean Time Between Routine Overhaul of an Engine	13500

The failure data collected from 70 engines are shown in Table 17.13. These failures have resulted in unscheduled maintenance of the engine.

**TABLE 17.13** In-service failure data of tank engine

215	114	247	122	91	291	194	315	241	9
95	378	425	252	22	153	195	165	130	451
456	350	105	275	294	232	441	94	360	202
18	50	68	244	126	557	168	64	99	323
42	560	232	198	27	947	239	89	325	273
178	33	997	75	126	23	416	247	292	216
210	36	46	223	284	52	223	454	151	88

The mean time between failures (which resulted in unscheduled maintenance) is 227.61. Since the target MTBF is 420 engine hours (Table 17.2), the manufacturer has to improve the MTBF since the achieved MTBF is much lower compared to the target MTBF. Defence procurement procedures (DPP) insist that the MTBF specified by the supplier at the time of procurement should be met to avoid any penalty,

### **Define stage**

In this case the CTQ is MTBF, thus the problem is to improve the MTBF.

### **Measure Stage**

From data in Table 17.13, estimated value of MTBF is 227.61. To understand which components are failing frequently and causes of failure, one has to deep dive into the data set to understand causes of low reliability in detail. Table 17.14 provides the engine strip data which identifies the parts that failed and the corresponding cause. The probability density function of time to failure random variable (using data in Table 17.13) is a Weibull distribution with scale parameter  $\eta = 246.361$  and shape parameter

$\beta = 1.286$ . The probability density function (pdf) of time to failure is shown in Figure 17.7. Since the shape parameter value  $\beta$  is greater than one, the armoured vehicle has increasing failure rate, that is the rate of failure increases with time.

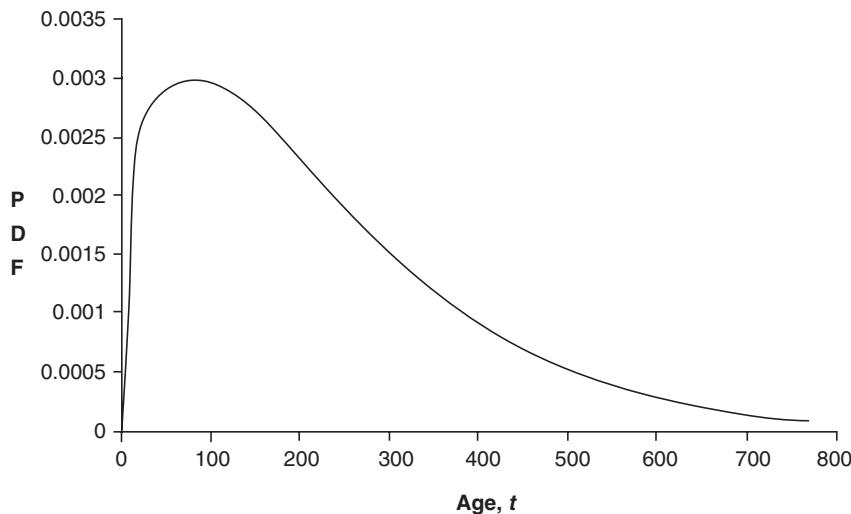


FIGURE 17.7 Probability density function of time to failure of Weibull random variable.

### Analyse Stage

In the measure stage, we identified the time-to-failure random variable as Weibull distribution, but it is important to understand the vital causes of failure of the engine. This can be achieved by analysing the data in Table 17.14 which provided micro-level data related to failure of parts within the engine and the corresponding cause of failures. The data in Table 17.14 is used to find most frequently failing parts and vital causes. These are shown in bar charts (Pareto charts) in Figures 17.8 and 17.9.

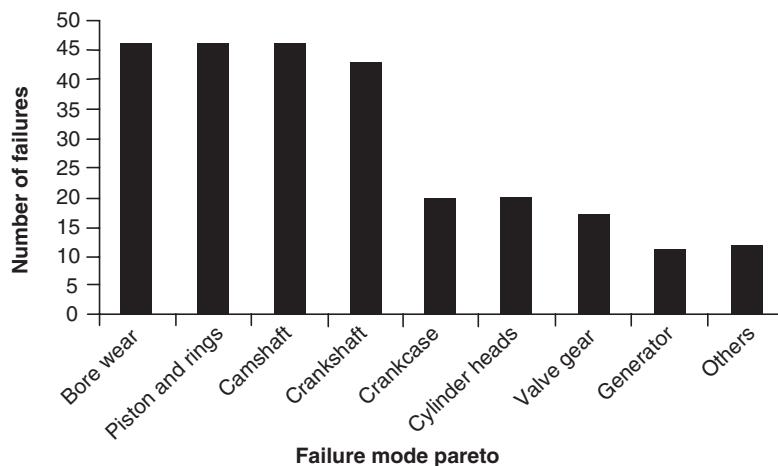
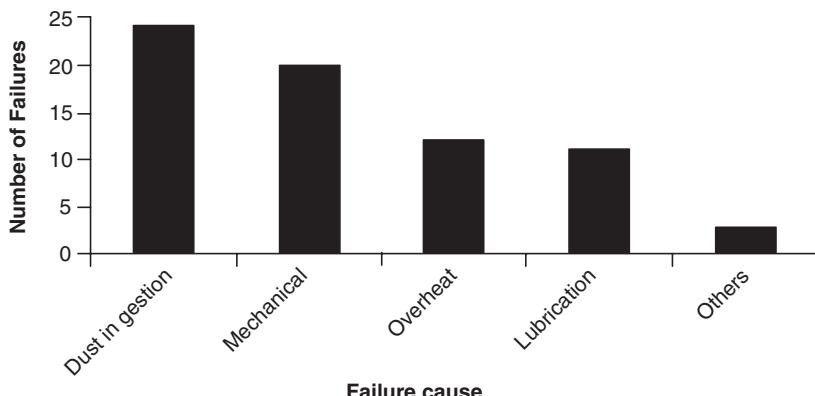


FIGURE 17.8 Frequently failing parts.

It is evident from Figure 17.8 that majority of the tank engine failures are caused by bore wear, piston and rings, camshaft and crankshaft. The major causes of failure are dust ingestion, mechanical failure (crack and corrosion), overheating and lubrication (Figure 17.9).



**FIGURE 17.9** Major failure causes.

### Improve Stage

We have identified the vital causes of failure in Analyse Stage. In the improve stage we develop ideas/solutions for causes identified. The following are few possible solutions:

1. Designing a better air filter to stop dust ingestions.
2. Mechanical failures such as crack and corrosion are caused due to water ingress which can be stopped by redesigning the engine cover to stop water ingress.
3. Use of internet of things and analytics to predict failures in advance and plan preventive maintenance.
4. Make necessary design changes to stop overheating of the engine.
5. Optimal lubrication of bearings to stop bearings-related failures.
6. It is also evident from Table 17.14 that there are many failures that are no fault found (NFF). NFF is a common problem in complex systems in which the user reports a fault; however at the repair workshop they do not find any fault. A better training across the support system will reduce NFF.

The above are few possible solutions for the vital causes identified, but different solutions will have different cost impacts. So it is important to find the solutions that will give maximum impact at minimum cost.

### Control Stage

This is a reliability improvement project and thus continuous monitoring of the reliability through frequent collection of data is required. Since design changes itself may be incremental, it has to be continuously monitored through data collection for validation of actual improvement.

**TABLE 17.14** Engine maintenance data – physical fault and cause

Strip Report Physical Fault												Cause of Failure					
Engine No	Time To Failure	Bore Wear	Pistons and rings	FIE Failure	Valve gear	Crankshaft or Bearings	No FF	Camshaft fault	Gen drive	Crank case	Fans	Cylinder Heads	Lubrication	Overheat	Mechanical	Dust Ingestion	No Fault Found
R4	215	0	0	0	1	1	0	1	1	0	0	0	0	0	1	0	0
L2	114	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0
M7	247	1	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0
F9	122	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0
M5	91	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0
L4	291	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0
I5	194	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
J8	315	1	1	0	1	1	0	0	0	0	0	0	0	0	0	1	0
W2	241	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0
Q7	9	1	1	0	0	1	0	1	0	0	0	0	1	0	0	0	0
W4	178	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0
E2	33	1	1	0	0	1	0	0	0	1	0	0	1	0	0	0	0
R3	997	1	1	0	1	1	0	0	0	0	0	0	0	0	1	0	0
T1	75	0	0	0	1	0	0	1	0	0	0	1	0	1	0	0	0
Y1	95	0	0	0	1	1	0	0	1	0	0	0	0	0	1	0	0
I2	378	1	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0
O2	425	0	1	0	0	1	0	0	0	1	0	0	1	0	0	0	0
P5	252	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0
A9	22	1	1	0	1	0	0	0	0	0	0	1	0	0	0	1	0
S7	153	1	1	0	1	1	0	0	0	1	0	1	0	1	0	0	0
D4	195	1	1	0	1	0	0	0	0	0	0	1	0	0	0	1	0
F8	165	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0
G9	130	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0
H2	451	1	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0
J3	126	1	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0
K5	23	1	1	0	1	1	0	1	0	0	0	1	0	0	1	0	0
L5	416	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0
Z7	247	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0
X9	456	1	1	0	0	1	0	0	0	1	0	0	0	0	0	1	0

(Continued)

**TABLE 17.14** Engine maintenance data – physical fault and cause—Continued

Strip Report Physical Fault													Cause of Failure				
Engine No	Time To Failure	Bore Wear	Pistons and rings	FIE Failure	Valve gear	Crankshaft or Bearings	No FF	Camshaft fault	Gen drive	Crank case	Fans	Cylinder Heads	Lubrication	Overheat	Mechanical	Dust Ingestion	No Fault Found
C8	350	1	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0
V6	105	1	1	0	1	1	0	1	0	0	0	0	0	1	0	0	0
B2	275	1	1	0	1	1	0	0	1	0	0	1	0	0	0	1	0
N1	294	1	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0
M4	232	1	1	0	0	1	0	0	0	1	0	0	1	0	0	0	0
U6	441	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
A1	94	1	1	1	0	1	0	0	0	0	0	0	0	0	1	0	0
S2	360	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0
D4	202	0	0	1	1	0	0	1	0	0	0	0	0	0	1	0	0
F3	210	0	0	0	0	1	0	0	0	0	1	1	0	1	0	0	0
G5	36	1	0	0	0	1	0	0	0	0	1	0	1	1	0	0	0
H6	46	0	0	1	0	1	0	0	1	0	0	0	0	0	1	0	0
J8	223	1	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0
K7	18	1	1	1	0	1	0	1	0	0	0	0	0	0	1	0	0
L9	50	1	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0
Q1	68	1	1	0	1	1	0	0	1	0	0	1	0	0	1	0	0
W2	244	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0
E3	126	1	1	1	1	0	0	0	0	0	0	1	0	0	0	1	0
R7	557	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0
T4	168	1	1	0	1	1	0	0	1	0	0	0	0	0	1	0	0
Y5	64	1	1	0	0	1	0	0	0	1	0	0	0	0	1	0	0
U6	99	0	0	0	0	1	0	0	1	0	1	1	0	1	0	0	0
I7	323	1	1	0	0	1	0	0	0	0	0	1	0	0	0	1	0
O8	284	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0
P9	52	1	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0
Q2	223	1	1	0	0	0	0	0	0	1	1	1	0	1	0	0	0
Z3	454	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0
Y4	42	1	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0
B5	560	1	1	0	0	1	0	0	0	1	0	0	1	0	0	0	0

**TABLE 17.14** Engine maintenance data – physical fault and cause—Continued

Strip Report Physical Fault													Cause of Failure				
Engine No	Time To Failure	Bore Wear	Pistons and rings	FIE Failure	Valve gear	Crankshaft or Bearings	No FF	Camshaft fault	Gen drive	Crank case	Fans	Cylinder Heads	Lubrication	Overheat	Mechanical	Dust Ingestion	No Fault Found
D6	232	1	1	0	0	1	0	0	0	0	0	1	0	1	0	0	0
S7	198	0	0	0	0	1	0	0	0	0	1	1	0	1	0	0	0
L8	27	1	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0
09	947	1	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0
J0	239	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1

## 17.15 | SIX SIGMA TOOLBOX

Six Sigma uses many statistical tools during different stages of DMAIC methodology to achieve the objective of different stages. Table 17.15 shows various tools that are used during different stages of Six Sigma. Note that Table 17.15 is only a sample set of tools used during various stages of DMAIC and is not an exhaustive set of tools. Note that few techniques are used across many stages of DMAIC. For example, quality function deployment can be used for identifying critical to quality (CTQ) during Define Stage and to find critical to process (CTP) during Analyse Stage.

**TABLE 17.15** Techniques used during various stages of DMAIC methodology

Define	Measure	Analyse	Improve	Control
1. Voice of Customers	1. DPMO, Yield and Sigma Score Calculation	1. Quality Function Deployment.	1. Theory of Inventive Problem Solving.	1. Statistical process control (control charts).
2. Cause and Effect Diagram	2. Pareto diagram.	2. Regression Models	2. Mathematical Programming Techniques.	2. Process Capability Analysis
3. Suppliers, Inputs, Process, Outputs and Customers (SIPOC)	3. Taguchi Loss Function	3. 5 Whys	3. Concept Selection	
4. Quality Function Deployment (QFD)	4. Cost of Poor Quality (CoPQ) analysis.	4. Analytic Hierarchy Process	4. Pugh Matrix	
5. Balanced Score Card (BSC)		5. Design of Experiments		
		6. Failure modes effects and criticality analysis (FMEA)		

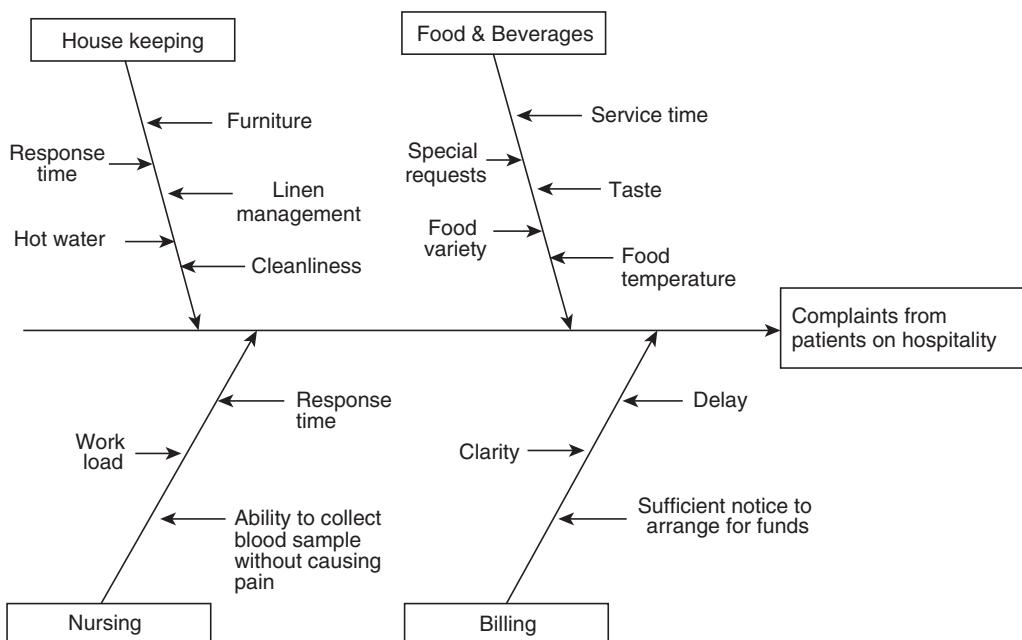
We will be discussing few of the techniques that are used in DMAIC methodology.

### 17.15.1 | Cause-and-Effect Diagram

Cause-and-Effect Diagram (also known as fish bone diagram or Ishikawa diagram) is a diagram created by Ishikawa (1968) that is used for identifying factors that contribute to an event. In the case of Six Sigma, we are interested in identifying the causes that contribute to defect as defined in the project scope. The causes are classified under different groups as listed below:

1. Cause category – 3Ms: Material, Machine, and Methods
  2. Cause Category – 4Ps: People, Policies, Procedures, and Plant

Each category can be further divided into subcategories. Note that the categories for a specific project need not be from all categories listed above and there could be new categories depending on the context. Categories form the branches that are connected to the backbone of the fishbone diagram. Figure 17.10 shows a cause-and-effect diagram in which the complaints are received from different departments of the hospital. In this case, complaints from patients about hospitality are the effect since each complaint is treated as a defect. Cause-and-effect diagrams are usually used during the design stage to find the sources of problems and CTQs.



**FIGURE 17.10** Cause-and-effect diagram.

17.15.2 | SIPOC

Suppliers, Inputs, Process, Output, and Customers (SIPOC) is a high-level process mapping technique used during Define and Measure stages of DMAIC methodology. SIPOC is useful for identifying CTQs and CTPs and process weakness that may be present. Following steps are used in SIPOC methodology:

1. Identify the process – process is a set of activities performed to convert inputs to the outputs.
2. Identify the outputs that come out of the process, outputs can be used for identifying CTQs.
3. Each output generated from a process should have a customer. This ensures the completeness of the process.
4. Identify the list of inputs to the process. Inputs can be used for identifying critical to process.
5. Identify the suppliers of the inputs.

An example of SIPOC for an automatic dosa (pan cake)-making machine is shown in Table 17.16.

**TABLE 17.16** SIPOC for automatic dosa-making machine

Suppliers	Inputs	Process	Output	Customers
Internal Suppliers of Batter & Accompaniments	Dosa Batter	Transfer right quantity of dosa batter to pre-heated tawa.	Dosa	Customer
	Oil	Add oil	Damaged Dosa	Waste management
	Accompaniments	Cook		
Power supplier		Spread accompaniments such as masala		
	Electricity	Remove and serve it to the customer	Steam	Exhaust for steam removal

SIPOC is a useful technique for identifying process weaknesses, CTQs, and CTPs. In the example described in Table 17.16, power supply is a process weakness due to frequent power cuts in India. So the design should have an option of backup power supply through batteries. Number of damaged dosa and number of dosas per hour are important CTQs.

### 17.15.3 | Five Whys

Five whys (or 5 whys) is an iterative technique for finding the root cause of a problem used during Analyse stage of DMAIC. 5 whys originated from Toyota corporation and logic 5 whys is that by asking why 5 times the vital cause of the problem is likely to be revealed. An example of 5 whys is shown in Table 17.17 in which the CTQ is complaints received in a hospital and one of them is about the quality of food delivered to a patient (food served was not hot).

**TABLE 17.17** Example of 5 whys

Why	The food served was not hot
Why	Food was delivered long time after it was prepared by the kitchen staff
Why	Kitchen staff did not inform the food delivery staff that the food was ready for delivery
Why	Kitchen staff assumed that the delivery staff will come and collect the food for delivery
Why	Lack of communication between kitchen staff and delivery staff

In the example discussed in Table 17.17, the CTQ is quality of food delivered to a patient and the root cause is the lack of communication between kitchen staff and delivery staff. There are a large number of tools used in Six Sigma including regression and linear programming discussed in Chapters 9, 10, and 15.

**SUMMARY**

1. Six Sigma is a powerful tool used for solving problems by improving the underlying processes.
2. Six Sigma mathematically means 2 defects out of one billion opportunities. For a Six Sigma process, the process capability index value will be 2.
3. Six Sigma sets the LSL and USL at  $\mu - 6\sigma$  and  $\mu + 6\sigma$ , respectively.
4. Defects per million opportunities, yield, Sigma Score are macro level metrics used to measure process performance in a Six Sigma process.
5. Six Sigma uses the framework Define, Measure, Analyze, Improve, and Control to improve existing process. DMAIC uses several techniques in its five stages to solve the problem.
6. In analytics, Six Sigma is an important tool since business process improvements are frequently carried out by organizations using Six Sigma.

**MULTIPLE CHOICE QUESTIONS**

1. LSL and USL in a Six Sigma process are
 

(a) $\mu - 3\sigma$ and $\mu + 3\sigma$	(b) $\mu + 3\sigma$ and $\mu - 3\sigma$
(c) $\mu - 6\sigma$ and $\mu + 6\sigma$	(d) $\mu + 6\sigma$ and $\mu - 6\sigma$
2. Proportion of defects in a Six Sigma process is
 

(a) 3.4 defects out of one million opportunities	(b) 2 defects out of one billion opportunities
(c) 3.4 defects out of one billion opportunities	(d) 6 defects out of one million opportunities
3. Sigma Score for a process with a yield of 99% is
 

(a) 3	(b) 4.5	(c) 2.32	(d) 6
-------	---------	----------	-------
4. Process capability of a Six Sigma process is
 

(a) 2	(b) 1	(c) 1.33	(d) 6
-------	-------	----------	-------
5. DMAIC methodology is used when
 

(a) Designing a new process and product	(b) Improving an existing process
(c) Improving an existing process and designing a new process	(d) Improving Sigma Score beyond 6
6. Sigma score is
 

(a) Area under normal distribution corresponding to the standard deviation of the process.	(b) Area under standard normal distribution corresponding DPMO.
(c) Area under standard normal distribution corresponding DPU.	(d) Area under standard normal distribution corresponding yield.
7. For a process
 

(a) Sigma and Sigma score should be high.	(b) Sigma and Sigma score should be low.
(c) Sigma should be low and Sigma score should be high.	(d) Sigma should be high and Sigma score should be low.
8. In DMAIC, the problems are defined using
 

(a) Critical to process parameters	(b) Sigma Score
(c) Critical to quality parameters	(d) Cost of poor quality

**EXERCISES**

- At Die Another Day (DAD) hospital, the target waiting time for registering new patients is 15 minutes. Any patient waiting for more than 15 minutes is treated as a defect. During the recent data analysis, the Sigma Score for waiting time was found to be 2.8. In a month, about 2000 patients are admitted in the hospital. How many of them waited for more than 15 minutes?
- A cookie manufacturer sells 100 gram cookie packets. Due to several reasons, the cookie weight may not be exactly 100 grams. If the LSL and USL are 95 and 105 grams, what should be the maximum standard deviation to achieve Six Sigma process capability? Assume that the mean weight of cookie packets is 100 grams.
- The time taken to discharge a patient in a hospital depends on many factors such as insurance clearance, clearing bills not covered under insurance, etc. A hospital sets a target that the time to discharge a patient should be less than 120 minutes; any discharge taking more than 120 minutes is considered as unacceptable. Analysis of previous data has shown that the average time taken to discharge a patient is 150 minutes with a standard deviation of 20 minutes (assume normal distribution). Calculate the Sigma Score of the discharge process. If the calculated Sigma Score is negative, explain under what conditions the process will have negative Sigma Score.
- A hospital classifies customers who give a feedback of less than or equal to 6 (on a 11 point scale, that is 0–10 scale) as detractors under the net promoter score (NPS) calculation. The hospital collects feedback on the following aspects:
  - Overall engagement
  - Discharge process
  - Nursing staff behaviour
  - Doctor attitude
  - Food and beverages

In the month of October 2016, there were 7750 patients treated by the hospital and the scores given by these patients on the aforementioned aspects are provided in the following Table 17.18.

**TABLE 17.18** NPS data

Department	Score $\leq$ 6	Score $>$ 6
Overall Engagement	28	7240
Discharge Process	72	7590
Nursing Staff Behaviour	36	7440
Doctor Attitude	12	7640
Food and Beverages	112	6800

Note that not all patients provided feedback and those who provided feedback may not have rated all the departments. If detractors are treated as defects, what is the current Sigma level? If the target is to achieve a Sigma level value of 3, what should be the maximum number of detractors every month, assuming that the number of patients admitted every month in the hospital will be 7750?

- A telephone service provider estimates that the percentage of dropped call rate is 5%. Calculate the corresponding Sigma Score. What should be the target dropped call rate if the telephone service provider would like to achieve a Sigma Score of 3.5?

## Era of Quality at the Akshaya Patra Foundation<sup>2</sup>

Quality is all about delivering a safe, tasty, nutritious, hot meal on time and every time.

— Muralidhar, Quality Head,  
The Akshaya Patra Foundation

It was 4 am on Saturday, January 10, 2015. The Vasanthapura (VK) Hill kitchen of The Akshaya Patra Foundation (TAPF) was filled with the enticing aromas of pulao<sup>3</sup> being cooked for the Midday Meal Programme. Steam-sterilized utensils were laid out in a row on the conveyer belt. Huge automated cauldrons turned the raw rice into pulao with freshly chopped vegetables and spices. Thermally insulated vehicles queued outside the kitchen waiting to load the utensils that contained food to be supplied to schools. Children in many government schools across Bangalore waited eagerly for this food as this was the only wholesome meal of the day for most of them.

At 7 am, Muralidhar, quality head at TAPF, visited VK Hill kitchen<sup>4</sup> as part of his routine to interact with various people at the workplace. With only one hour remaining to dispatch the food to the schools, his thoughts automatically drifted toward the quality department's recent attempts to reduce the cooking cycle time. The number of schools served by VK Hill kitchen had increased steadily in the recent past resulting in longer cooking times, leading to delays in carrying the food to some of the faraway schools. He knew that these delays would only increase further as TAPF was planning to increase the number of children fed every day to 5 million from the current operational scale of 1.4 million. He was very keen to fix these delays before more schools were added to the VK Hill kitchen. TAPF was using two important key process indicators (KPIs): temperature of food at the time of delivery and number of deliveries made before 12.00 noon every day. Ensuring an acceptable level of these two KPIs and expanding TAPF's service to reach more schools were the primary challenges faced by the team at TAPF.

There were many Six Sigma projects that the quality team had initiated since 2012 in TAPF; one of them in particular focused on cooking cycle time reduction. Although many findings and potential solutions were suggested by the team members, limited funds were available for capital investments, making implementation of any suggestions that required large capital, unviable. The team reflected on whether there were any other means to reduce cooking cycle time, using Muralidhar's vast experience from the manufacturing sectors.

<sup>2</sup> © The Indian Institute of Management Bangalore. The case was co-authored by Srujana H M, Haritha Saranga, and U Dinesh Kumar. This case is not intended to serve as an endorsement or source of primary data, or to show effective or inefficient handling of decision or business processes. Reproduced with the permission of IIM Bangalore.

<sup>3</sup> Pulao or pilaf is a rice-based dish which often also includes ingredients such as vegetables and pulses. Pulao was one of the special rice dishes that was typically served on Saturdays, instead of the regular rice and Sambar that were served on week days.

<sup>4</sup> VK Hill kitchen: TAPF kitchen at Vasanthapura Hill area in Bangalore was established in July 2007 and is a centralized kitchen.

**Case Study** **Continued...**

## THE AKSHAYA PATRA FOUNDATION

### Inception

TAPF started with a story of compassion. Looking out of a window one day in Mayapur (a village near Calcutta, now Kolkata, India), Swami Prabhupada, founder-Acharya of the International Society for Krishna Consciousness (ISKCON), saw a group of children fighting with street dogs for leftover food. From this simple yet heartbreaking incident was born a determination to feed hungry children. Swami Prabhupada immediately affirmed that, '*No child within a 10-mile radius of ISKCON center should go hungry*'. It is his inspiration that helped to create TAPF in 2000.

As per published statistics of the United Nations Children's Fund (UNICEF), one in every three malnourished children in the world lived in India in the year 2000.<sup>5</sup> Malnutrition stifled development and the capacity to learn in children. As per the Indian Ministry of Labour and Employment, there were more than 10 million children engaged in child labour in the year 2000.<sup>6</sup> These statistics indicated that a sizable number of children who should have been in school were instead out on the streets trying to supplement their family income. In response to the United Nations Organisation's calls to end poverty and hunger and provide universal education,<sup>7</sup> the Supreme Court of India passed an order on 28 November 2001,<sup>8</sup> which mandated that cooked midday meals were to be provided in all the government and government-aided primary schools of all the states.<sup>5</sup> Inconsistent food quality, occasional food poisoning, poor hygiene, and operational concerns were some of the major challenges faced during the provision of government-sponsored midday meals.

In Karnataka, only dry ration (consisting of rice and pulses) was distributed to the parents of the government school children in lieu of the midday meal. However, anecdotal evidence suggested that this provision failed to ensure that the children received one square meal a day, as many fathers sold off the dry ration to cover their expenditure on alcohol. According to Venkatachallaiah, Assistant Headmaster of Government High School, Peenya, located on the outskirts of Bangalore:

There were incidents of children falling unconscious during the assembly, as many of them would come without eating breakfast or even dinner the previous night.

Moved by the plight of the children in government schools, missionaries of the ISKCON center in Bangalore, a few corporate professionals, and a few entrepreneurs came together to create TAPF. The Foundation's vision was, 'No child in India shall be deprived of education because of hunger', thereby addressing the twin challenges of hunger and education. TAPF carried out the government's Midday Meal Programme through the collaborative approach of Public-Private Partnership (PPP),

<sup>5</sup> Source: [http://www.unicef.org/india/children\\_2356.htm](http://www.unicef.org/india/children_2356.htm), accessed on May 29, 2014.

<sup>6</sup> Source: <http://labour.nic.in/content/division/child-labour.php>, accessed on May 29, 2014.

<sup>7</sup> Source: <http://www.un.org/millenniumgoals/>, accessed on June 2, 2014.

<sup>8</sup> Source: <http://www.sccommissioners.org/FoodSchemes/MDMS.html>, accessed on June 2, 2014.

**Continued...**

wherein around 51% of funding was provided by central and state government grants and subsidies, allowing the organization to focus on raising the rest of the funds required to operate the kitchens (**Exhibit 1**).

In the beginning, the Foundation provided midday meals to 1,500 children in five government schools in Bangalore. With the help of an efficient management committee, the Foundation grew and reached out to 14 lakh (1.4 million) needy children per day across many states of India by 2014. One of the schools to be initially selected by TAPF for distribution of midday meals was the Government High School in Peenya, as the founders discovered that the schools on the city's outskirts and rural areas were in maximum need of the Midday Meal Programme. Reflecting back, Venkatachallaiah commented

From the moment Akshaya Patra started providing mid-day meals, we found the children to be very energetic as they were assured of at least one nutritious meal per day.

Similar sentiments were shared by teachers and students across many government schools that were recipients of the midday meals from TAPF.

## Operating Model

The TAPF operating model involved setting up cooking infrastructure (kitchen) in a region that could cater to the demands of a number of government schools in and around that area, using delivery vans.<sup>9</sup> The capacity of the kitchen and the size of the delivery fleet were determined on the basis of the estimated demand within a region.<sup>10</sup>

## The Kitchens

Since its inception in 2000, TAPF has constantly evolved its kitchens from manual to automated and centralized to decentralized model.

### Centralized Kitchens

The centralized model was one where all the cooking activities of a particular location were concentrated within one large kitchen. Each centralized kitchen had the capacity to cook between 50,000 and 1,50,000 meals a day. Food was cooked in large quantities in centralized kitchens with a high degree of automation (**Exhibit 2**) and distributed to individual schools based on the number of students. TAPF operated two centralized kitchens in the city of Bangalore (one of them being the VK Hill

<sup>9</sup> Vans that carry the food to the schools are thermally insulated to ensure that food is delivered hot in a safe and hygienic manner.

<sup>10</sup> OR at Work in Feeding Hungry School Children, *Interfaces*, 43(6), 530–546.

**Continued...**

kitchen) to provide midday meals to 1,85,000 children in and around the Bangalore metropolitan area.

Case Study

### **Decentralized Kitchens**

TAPF expanded its Midday Meal Programme to rural areas of India that suffered from the maximum levels of poverty and hence have the most malnourished children. The *cooking-to-consumption time* (which should not exceed 6 hours as per quality standards) became one of the major reasons for this expansion. Transportation of cooked food to the schools in rural areas from centralized kitchens would take more time owing to lack of road infrastructure in rural India. Moreover, setting up a large kitchen facility was tricky owing to dispersed locations and difficult terrain of villages in the rural districts. Therefore, a decentralized model was used, which meant that TAPF needed a large number of smaller kitchens located close to the village schools. Decentralized kitchens posed an additional challenge of finding chefs and support staff who were willing to live in the villages. Decentralized kitchens of Rajasthan and Baran catered to 166 schools, while those of Orissa and Nayagarh catered to 352 schools.

By 2013, TAPF managed to set up 18 centralized and 2 decentralized kitchens across 9 states. It also meticulously set the standard workflow mechanisms in all its centralized kitchens (**Exhibit 2**).

## **BUILDING THE CULTURE OF QUALITY**

Quality is not only about the people and their work in the Quality Department. It is about the culture of the organization and it certainly is everybody's business

— Shridhar Venkat — CEO

As Akshaya Patra expanded its operations, it faced numerous quality challenges. For example, centralized and decentralized kitchens had different raw material procurement mechanisms. Some of the raw materials were procured through PPP, which further complicated the systems as there were numerous stakeholders. In addition, recipes and food preparation techniques varied significantly between the locations owing to local preferences. These complexities created quality issues owing to increasing variability in operations and management.

Some of the challenges faced by the management during the initial phase were: (a) how to standardize the processes; (b) ensuring that suppliers conform to the acceptable quality standards; and (c) ensuring safety in food preparation and delivery processes. TAPF found that quality accreditations were a good way to establish standards. In 2007, TAPF applied for the International Organization for Standardization (ISO) certification and six of its kitchens were certified by the end of 2008. Also TAPF received the Det Norske Veritas (DNV) certificate for food safety. By 2013, 11 out of the 20 kitchens of TAPF were ISO 22000 certified. TAPF also included inputs from the ISO 9001 Quality Management System to formulate its in-house best practices and good manufacturing practices (GMP).

Case Study  
Continued...

## Supplier Quality Management

Since its inception, TAPF procured rice, pulses, vegetables, edible oil, and spices from different suppliers. From 2001 onwards, the Food Corporation of India (FCI) provided 50% of the TAPF's rice requirement through subsidized raw rice (since TAPF was implementing the Midday Meal Programme mandated by the Supreme Court). TAPF mandated procurement of the best quality of raw materials for all kitchens and implemented a robust Supplier Quality Management System (SQMS). The SQMS process covered sub-processes such as supplier selection, supplier qualification, and supplier rating to ensure that the best raw materials were procured. The TAPF Quality Control (QC) process ensured that raw materials were accepted only after thorough Quality Inspection (QI). An example of raw material specifications is described in **Exhibit 3** for QI of potatoes.

## Quality during Precooking Preparations

Quality is about making children relish our meal every day

— Saanil Bhaskaran, *iGiving & Donor Care*

Although fresh vegetables were procured on a daily basis, proper washing and cutting of the vegetables to prepare them for the cooking process was a major challenge, given the scale of operations in centralized kitchens. Sanitization standards with a three-step process of cleaning vegetables were introduced. Washing the vegetables in chlorinated water with chlorine levels of  $75 \text{ ppm} \pm 20 \text{ ppm}$  was followed as a critical control point (CCP) in preprocessing. Vegetables were cut manually by the kitchen workers and, in their hurry to process them, they tended to cut them into bigger sizes. There were complaints from schools that children did not eat most of the vegetables. The vegetable pieces were too big and it was not easy for the children to chew and swallow them. In response, TAPF introduced automated cutting machines which could cut the vegetables into small and equal sizes very quickly. This introduction not only increased the vegetable consumption by children but also reduced the cutting time. Another automation in the kitchens was the addition of a roti-making machine, which was capable of producing 40,000 rotis per hour,<sup>11</sup> again bringing the cooking times under control, despite the increase in volume.

Another innovative idea was to separate precooking activities such as washing and cutting of vegetables from the actual cooking process through use of cold storage in centralized kitchens. This essentially allows the vegetables to be cut in the evening of the previous day, to store them in cold storage, and use them during the actual preparation of meals the next day. This makes the vegetable cutter's life easy as they could work in the evening and complete their jobs, rather than work at midnight. The temperature in the cold storage was maintained at  $0\text{--}4^\circ\text{C}$ , which was another CCP in preprocessing. To ensure all the raw materials were fresh, the kitchens followed

<sup>11</sup> Before the introduction of the roti-making machine, rotis were made manually one by one, which is a highly time- and labor-consuming activity.

**Continued...**

the FIFO (First In, First Out) and FEFO (First Expiry, First Out) methods while issuing the raw material for production. By doing so, the kitchens were able to properly identify, store, and retrieve raw materials in an appropriate manner.

### **Quality during Cooking Operations**

One of the major challenges that TAPF faced during its 14 years of evolution was balancing nutrition and taste simultaneously. According to Ajay Kavishwar, GM, Branding & Media, one of the perennial questions was: How can we achieve adequate nutrition and sensory appeal and create excitement in children about the meal? The additional challenge was to customize the meal to the local preferences and tastes of the children. He recalled an incident during the inception years of TAPF, when soya chunks were introduced in the meals in the VK Hill kitchen in an attempt to improve the nutrition value of the meal.<sup>12</sup> To their dismay, the kitchen officials found out after a few days that most children had refused to eat the dish, resulting in much wastage and the primary nutritional objective also not being met. An inquiry into the matter revealed that, based on the colour and texture of the cooked soya chunks, children had assumed soya chunks to be non-vegetarian and hence stopped eating it.<sup>13</sup> Thereafter, TAPF began to tackle such situations by creating awareness among children and their parents on the origin and importance of various food items being served, including soya chunks. The protein and energy values of any given meal were tested in an approved lab at TAPF and 100% adherence to recipes was ensured through well-trained cooks and quality checks by production supervisors.

In addition, TAPF started using vessels made of the best available grades (e.g., grade 304) of stainless steel for cooking and packing from 2008 onwards. Contamination was prevented by steam sterilization of vessels before cooking and before packing and by maintaining a temperature of 95°C during the cooking process.

### **Feedback from the Schools**

Good for health, great in taste to the children, which brings healthy smiles ... is quality  
*- Ajay Kavishwar – Branding & Media*

TAPF finally completed the circle of quality by connecting the two ends of the supply chain through a feedback mechanism that traced back the problems faced by the consumers to TAPF's internal processes and all the way to the suppliers. For example, during a summer afternoon in 2010, one of

<sup>12</sup> Source: Nutritive Value of Indian Foods, National Institute of Nutrition, Hyderabad: <http://ninindia.org/DietaryguidelinesforIndians-Finaldraft.pdf>, accessed on June 2, 2014.

<sup>13</sup> Note that India is one of the lowest meat consumers in the world, with 20–42% of the Indian population being vegetarian. Source: Food and Agriculture Organization of the United Nations, and United States Department of Agriculture Survey (2007): [http://en.wikipedia.org/wiki/Vegetarianism\\_by\\_country](http://en.wikipedia.org/wiki/Vegetarianism_by_country), accessed on June 3, 2014.

**Continued...**

the supervisors observed that curd was not being consumed at the schools and conveyed this to the quality department. A team visited the schools for investigation and discovered that the curd was too sour for the children's taste. The quality team held a meeting with the supply chain management department and it was decided that a replacement of the supplier was necessary. The quality team later found to their satisfaction that the curd consumption had increased significantly after a new supplier was introduced.

An internal customer care centre was set up to facilitate grievance redressal. Children were encouraged to call a toll-free number to share their opinions and, more importantly, to lodge complaints if any. The quality personnel in the kitchens reviewed the feedback and triggered appropriate corrective or preventive actions to be taken within 48 hours of a complaint being registered at the call centre. In January 2014, for example, an oral complaint was received from Basaweshwara Girls' High School, Bangalore, through the call centre that a worm had been found in the milk. The quality team immediately did a root cause analysis and determined that the worm had come from the tap of the vessel that was used to serve milk to the students at the school. The quality personnel advised the schools to clean the vessels thoroughly before using them. In a similar vein, TAPF institutionalized Good Manufacturing Process (GMP) audits and surprise audits to ensure food safety and quality, and measurement of key processes and systems performance through monthly reviews of relevant quality metrics.

These efforts resulted in significant improvements in the delivery of quality food to children in government schools by TAPF. In 2011, a survey conducted by AC Nielsen<sup>14</sup> found that in schools where TAPF's Midday Meal Programme was being implemented, the enrolment and attendance of the children had increased, their concentration levels and nutrition levels had increased, and the children were picking up various life skills. The survey also found that TAPF had managed to empower the women in nearby areas by providing employment in Akshaya Patra kitchens.

## THE ERA OF LEAN SIX SIGMA AT TAPF

Since its inception, TAPF had considered quality assurance as a core concern of all its activities. TAPF had proactively adopted numerous measures to ensure that quality standards were designed and met, whether it involved obtaining accreditation from ISO, or design of food safety mandates, or formation of a dedicated quality team in-house. In 2012, as a part of its strategic decision-making, TAPF recruited Muralidhar to head a Quality cell and with this started a new era of *lean Six Sigma* at TAPF. Muralidhar joined Akshaya Patra with an explicit mandate for introducing process excellence in all the kitchens. Muralidhar was a trained Six Sigma Black Belt from Motorola and had vast experience in the manufacturing sector. His first impression about TAPF's meal preparations was that it was 'no different' from a manufacturing environment with its massive volumes and complex set of operations. He was also impressed by the modern equipment, the automated facilities, and the quality standards followed in the centralized kitchens. Although processes were already present in every part of the organization and were standardized in functional areas such

<sup>14</sup> Source: <http://www.akshayapatra.org/ac-nielson-case-study>, accessed on May 29, 2014.

## Case Study

### Continued...

as finance and purchasing and within each region, he found that there was significant scope for further standardization in areas pertaining to food safety and quality.

As a first step towards delivering his mandate, which required him to measure and improve the current process performance across various kitchens, Muralidhar introduced a small set of metrics covering all operational processes (**Exhibit 4**). This exercise not only afforded him a good understanding of all critical processes ranging from material procurement, storage, food preparation, delivery, and customer satisfaction, but also whetted his appetite by revealing opportunities for further improvement. Given his training and background, his first choice naturally was to opt for an organization-wide Six Sigma implementation to bring about improvement. However, the challenges involved in adoption of Six Sigma as an improvement tool were significant, owing to very low education levels of the unskilled workers on the kitchen floors. Muralidhar recalled

When I looked at the minimum education levels in order to start the Green Belt training programs, guess what I found? Many kitchen workers had never even been to a school. Now tell me, how do you train someone with no knowledge of reading or writing, in Six Sigma tools? In my experience, even someone with a Bachelor's degree finds it hard to follow the statistical concepts of Six Sigma tools.

So, after a careful consideration of the situation at hand, Muralidhar came up with three different types of continuous improvement initiatives for the three categories of people involved in TAPF operations.

1. Kitchen workers were involved in Kaizen projects.
2. Office staff members were involved in continuous improvement projects such as Plan, Do, Check, and Act (PDCA).
3. Office staff members were trained on quality aspects and were equipped with various statistical tools to work on Lean Six Sigma projects using the Define, Measure, Analyse, Improve, and Control (DMAIC) methodology.

Kaizen activities essentially sought any ideas for improvement, based on the working knowledge of kitchen workers pertaining to their areas. To encourage participation of workers in Kaizen, TAPF instituted a cash award ranging from INR 50 to INR 100 (\$1 ≈ INR 61, in January 2014) and quarterly Kaizen Thunder awards (see **Exhibit 5** for an award winning safety-related Kaizen idea in September 2013). Continuous improvement projects such as PDCA, on the other hand, involved slightly more educated workers and office staff, who could come up with improvement plans, carry out some amount of experimentation, monitor the progress, and make appropriate changes. The Six Sigma projects were more broad-based in scope and required proper planning by trained and dedicated personnel (project leader, champion, team members, etc.) and sufficient resources (from deployment champion), and required anywhere from three to four months for completion. Fourteen Six Sigma projects were pursued and completed in 2013 and six PDCA projects were pursued in 2014.

**Continued...**

Despite the great strides achieved by Akshaya Patra in ensuring quality across its value chain, the increasing number of schools/children being added to the Midday Meal Programme gave rise to more challenges every day. With a desire to serve as many children as possible, TAPF added more and more schools to the distribution networks of existing kitchens. The total strength of beneficiaries fed by VK Hill kitchen every day across 650 schools was 87,045 children. Based on the consumption pattern of the schools, on average, 7 tonnes of rice and 16,800 litres of Sambar (a spicy South Indian dish of lentils and vegetables) had to be cooked and dispatched by VK Hill kitchen every day. Based on the number of cauldrons being used and their cooking capacity, 0.875 tonnes of rice and 8,400 litres of Sambar were cooked per batch.

Although every additional child meant a decreasing marginal cost in case of a centralized kitchen (until the number reached its full capacity), the cooking times increased in proportion to the batches being cooked. This meant that the cooking needed to start at much earlier times, but there was an upper limit to the cooking-to-consumption time;<sup>15</sup> it could not exceed normally 6 hours. Food Safety and Standards Authority of India (FSSAI) mandated that cooked food served hot should be kept at a temperature of at least 60°C. Based on its logistics parameters, TAPF realized that ‘cooking-to-consumption’ time of 6 hours was the ideal duration to maintain the temperature specified by FSSAI. The lunch break in schools was between 12 noon and 1 pm, which meant that the delivery vans had to reach the schools by 12.00 noon. The farthest school in VK Hill’s serving area could be reached in 4 hours (43 km). The total packaging and loading into each of the 35 delivery vans took approximately 30 minutes. This brought the focus squarely on the cooking cycle time.

The cooking cycle time, therefore, was selected as one of the *critical-to-quality* (CTQ) measures and a Six Sigma project was initiated to improve this measure at VK Hill kitchen. With the help of the learning from the three-day Green Belt training program that the team had undergone, it was decided that the DMAIC methodology should be used to carry out this project. For most of the team members (other than Jaya Kumar – Manager CI Programs) this was their first Six Sigma project.

## Define

The team defined the problem as follows:

For the last 6 months, the total production process, which included 4.30 hours of preprocessing and vessel sterilization, took on average 8.30 hours. This resulted in occasional late delivery of the finished product to the distribution team.

The goal was: *To reduce the cooking cycle time by 1 hour in VK Hill kitchen in the next 3 months.*

The team then had a detailed look at the process using the Suppliers, Inputs, Process, Outputs, and Customers (SIPOC)<sup>16</sup> methodology and determined that given the nature of operations at TAPF, it was the gaps in inputs and processes which needed to be targeted.

<sup>15</sup> It is measured as the time from when cooking is completed until the consumption takes place.

<sup>16</sup> SIPOC is a process excellence tracing methodology that is used to determine the sequence and effectiveness of a process in a structured way. Source: <http://ianjseath.files.wordpress.com/2009/04/sipoc.pdf>, accessed on June 1, 2014.

## Case Study Continued...

### Measure

During the measurement phase, the team collected data on 106 samples. The average production time based on these samples was found to be 500 minutes, which was in line with the initial observations.

### Case Study

### Analyse

Next, the team decided to analyse the data they had collected to identify the root causes for longer cooking cycles. Rice and Sambar were the major food items that were cooked; therefore, the team identified various activities involved in cooking these two items, created process maps, and computed the respective process times for each of these activities (**Exhibits 6 and 7**).

The description of various activities in the process maps was helpful in identifying why it took longer to cook the items (especially Sambar), and whether there was any scope for improvement. The team members also made several visits to the kitchen to observe the cooking and packaging activities to identify causes for any delays. These visits, interviews with the kitchen staff, and many brainstorming sessions within the team members (which also included some of the kitchen staff members) using the '5 Whys' analysis helped them to identify the root causes (which are listed in the Ishikawa Fish Bone diagram in **Exhibit 8**). For example, one of the kitchen staff members complained that whenever the packaging specification was to fill up only half the vessel, it took longer than filling up the full vessel. Such counter-intuitive factors led the team to believe that there was a need to collect further data on certain activities and carry out further analysis.

The primary root causes were then segregated with the help of Pareto analysis,<sup>17</sup> which was used to identify few vital root causes, whose elimination or improvement would lead to significant improvement in the cooking cycle time (**Exhibit 9**). They subsequently gathered more data on these primary root causes. The data were collected on packaging specifications and numbers of vessels with their respective specifications. Although non-availability of vessels was identified as one of the potential causes for delays in cooking during the brainstorming sessions by the team members, the data collection did not show any incidents where shortage of vessels was leading to high cooking cycle times. Therefore, it was concluded that shortage of vessels was not really a reason for longer cooking cycle times.

### WHAT NEXT?

Muralidhar contemplated on how best to align the quality improvement program with the massive expansion plans that TAPF was drawing up to benefit more needy children. As organizational processes were scaled up to meet the expansion plans, there was a need for continuous improvement

<sup>17</sup> Although Pareto analysis identifies vital root causes based on their frequency of occurrence, owing to lack of data on this front, the team decided to brainstorm and use voting method to identify the important causes, thereby bringing in the experience of team members to fill this gap in data.

**Continued...**

programs to identify new improvement opportunities and exploit them. Muralidhar felt that too much time and money was being spent on firefighting, while the mission of feeding five million children in the near future required all organizational processes to be more efficient and effective.

Primary analysis by the Six Sigma team had suggested that if the cooking operational efficiencies were improved, there was a possibility of reaching out to more needy children without a substantial increase in costs. As the cooking-to-consumption time was one of the critical control points for the Midday Meal Programme, it had to be tackled not only to enhance cooking operational efficiency, but also to ensure that the quality standards are met. Cooking cycle time reduction was one of the first Six Sigma projects initiated by Muralidhar and his team and, he was hoping that success in this project – in terms of efficiency gains if not cost savings – would boost the morale of his team members and encourage them to take up other critical projects with increased vigour.

### **Exhibit 1: Akshaya Patra Funding Details**

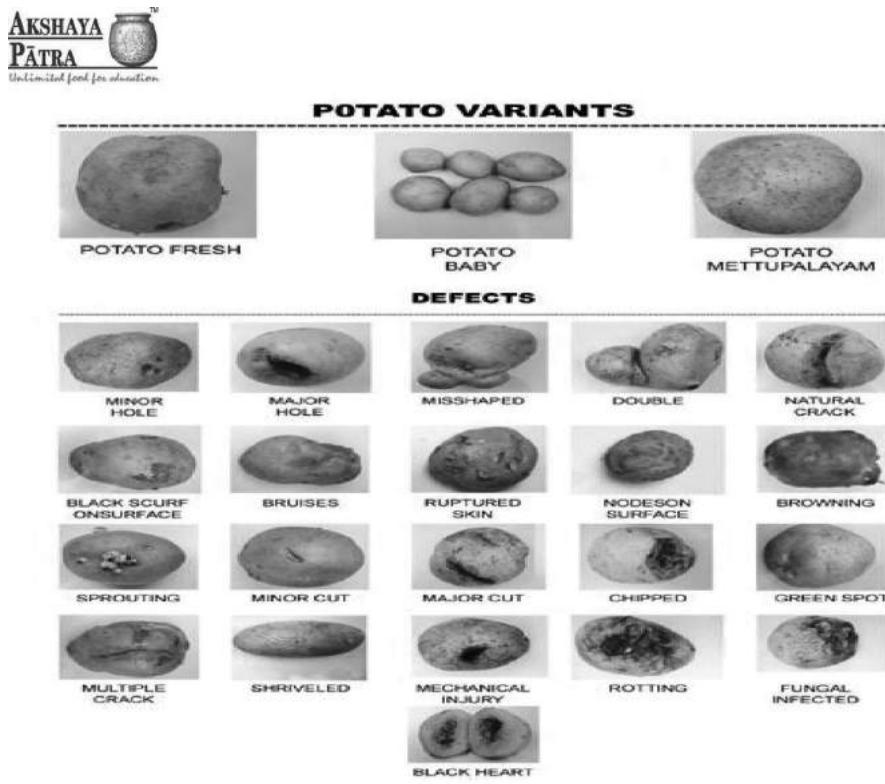
TAPF carries out the Midday Meal Programme through the collaborative approach of PPP. As of March 2013, around 51% of funding (amounting to Rs. 40,668 lakh) has been accounted for by central and state government grants and subsidies, leaving the organization to focus on raising the rest. This arrangement entrusts TAPF with the responsibility of fund generation for sustainability of its operations. Despite the government contribution, TAPF has to bear the partial expense of feeding the children that accounts to approximately Rs. 750 per child per year. Thus, if TAPF intends to expand its operations, the efforts toward fund generation also need to be increased equally. As of 2014, TAPF was able to generate funds through generous corporate and private sponsorships; however, there was no guarantee that the donations would either continue or grow year on year. Moreover, expenses were likely to increase year on year owing to market inflation which would further intensify if TAPF were to expand its operations. Therefore, in order to generate additional revenue for its core charitable operation of the Midday Meal Programme, the management committee proposed entry into other segments by leveraging its unused capacity and competencies through another charitable trust called Akshaya Nidhi. Akshaya Nidhi intends to donate its entire disposable surplus (profit after tax minus amount retained for business needs) for charities. It intends to lease out the unutilized production capacities of Akshaya Patra kitchens to carry out the business of production and sale of high-quality cooked food for the masses at affordable prices. Akshaya Nidhi also intends to leverage the strengths of Akshaya Patra in the areas of sourcing, distribution, production, and quality, which TAPF has developed over a period of 13 years, to help generate funds that would ensure sustainability of the Midday Meal Programme.

Case Study •

Continued...



**EXHIBIT 2** Workflow in centralized kitchens. Source: The Akshaya Patra Foundation.



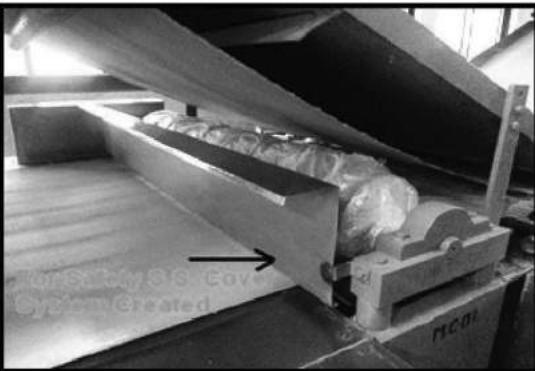
**EXHIBIT 3** Raw material specifications. Source: The Akshaya Patra Foundation.

Continued...

**EXHIBIT 4** Quality metrics

Supplier Quality Management	Lot Acceptance % Repeat Issues Rate
Stores and Material Handling Quality Metrics	Stores Best Practices Index
Cooking Quality Metrics	Food Quality Index Production Process Compliance %
Food Transportation Quality Metrics	Vehicle Hygiene Index On Time Delivery %
Customer Satisfaction Quality Metrics	Customer (School) Complaints Customer Satisfaction

Source: The Akshaya Patra Foundation.

 KAIZEN THUNDER		SI #: TAPF - KT- VDD - 059 Location : VADODARA Dept MAINTENANCE DEPT.
		<b>The likelihood of an accident.</b>
DESCRIPTION : In chapatti machine there was no safety barricade to safe guard the workers hand near chapatti cutting die	KAIZEN DESCRIPTION: Installed S.S. Guard before die roller to safeguard the hands. ADVANTAGES/VALUE ADD: Prevent accidents.	
IMPROVEMENT CATEGORY (QUALITY/COST/CYCLE TIME/SAFETY/MORALE) : Safety		
APPROVALS: KAIZEN IDEA BY : GANESH ( ROTI OPERATOR ) DATE: 14 Sept 2013 IMPLEMENTED BY : AJAY WAGH ( MAINTENANCE INCH ) DATE: 20 Sept 2013 APPROVED BY : TUSHAR DANGE ( Sr. Opr's Manager ) DATE: 20 Sept 2013	VERIFICATION DESCRIPTION: ( Check after 10 DAYS ) Sustained VERIFY BY : SUDHANSU TRIPATHI ( Q.C. ) DATE: 30 Sept 2013 APPROVED BY : TUSHAR DANGE ( OPR'S MNGR ) DATE: 30 Sept 2013	

**EXHIBIT 5** An example of a Kaizen initiative. Source: The Akshaya Patra Foundation.

## Case Study

Continued...

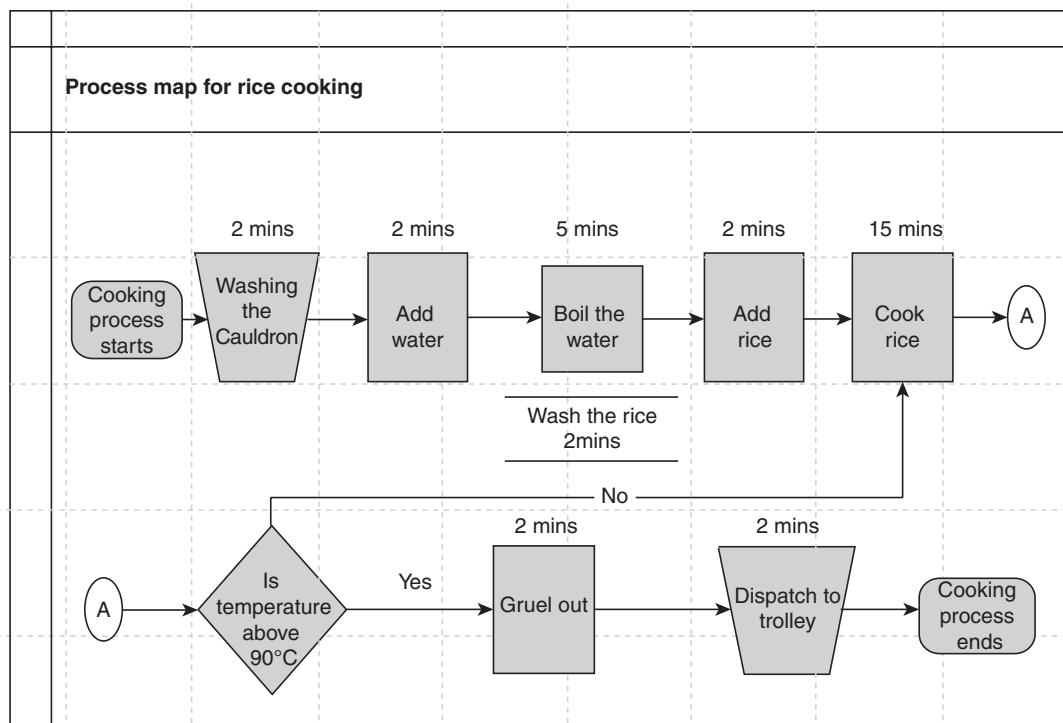


EXHIBIT 6 Process map for rice cooking. Source: The Akshaya Patra Foundation.

Case Study

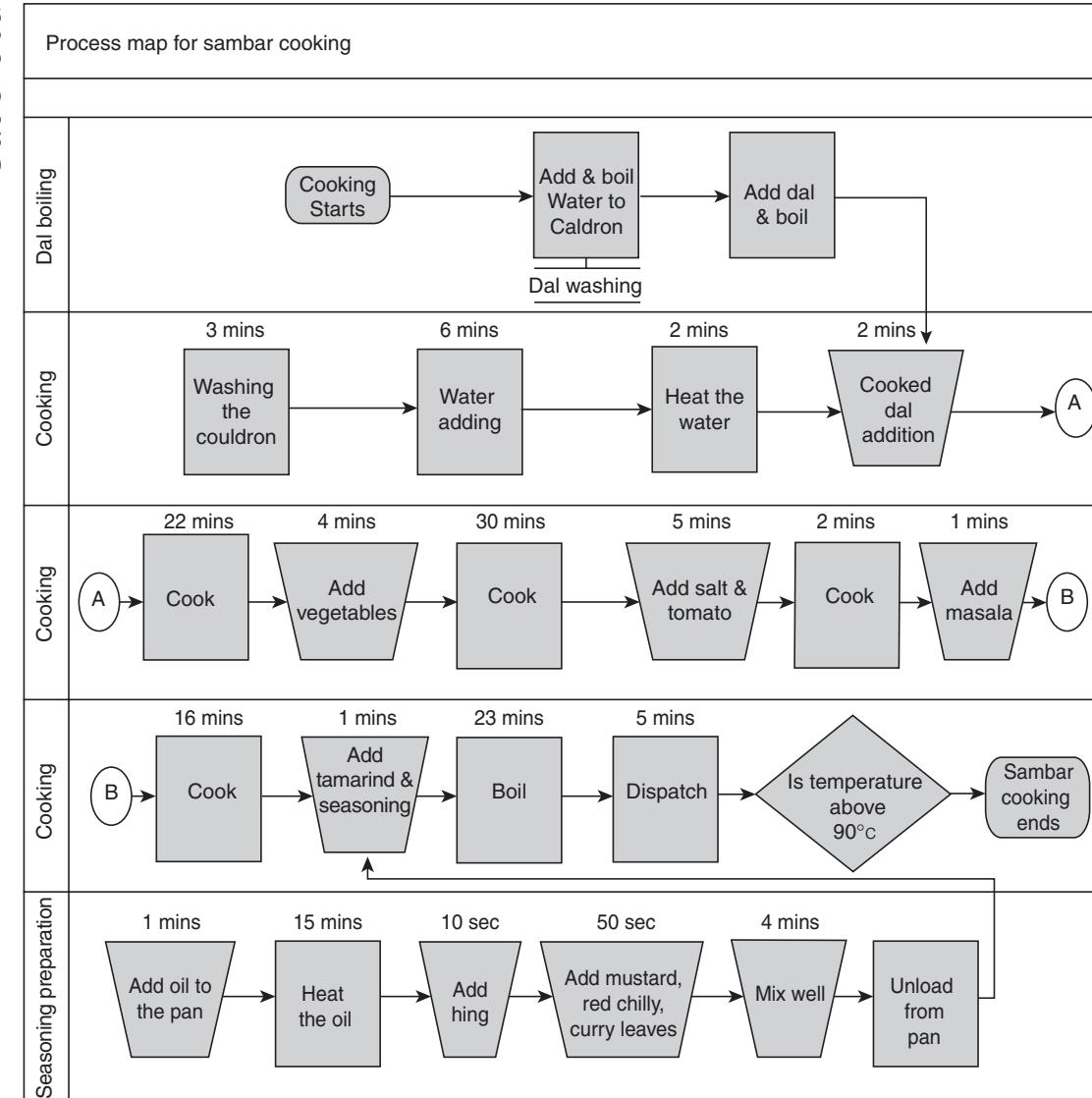
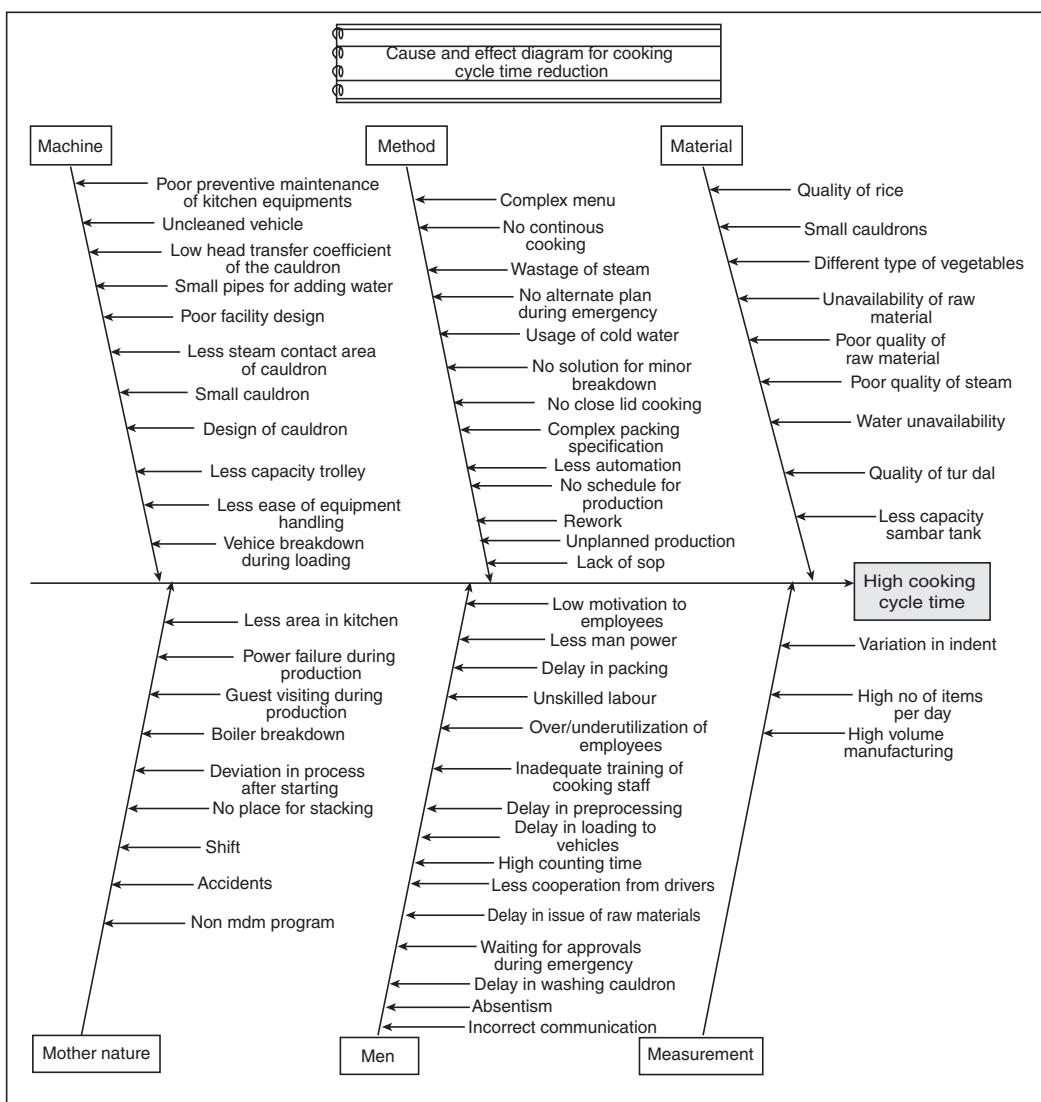


EXHIBIT 7 Process map for sambar cooking.

Case Study

Continued...



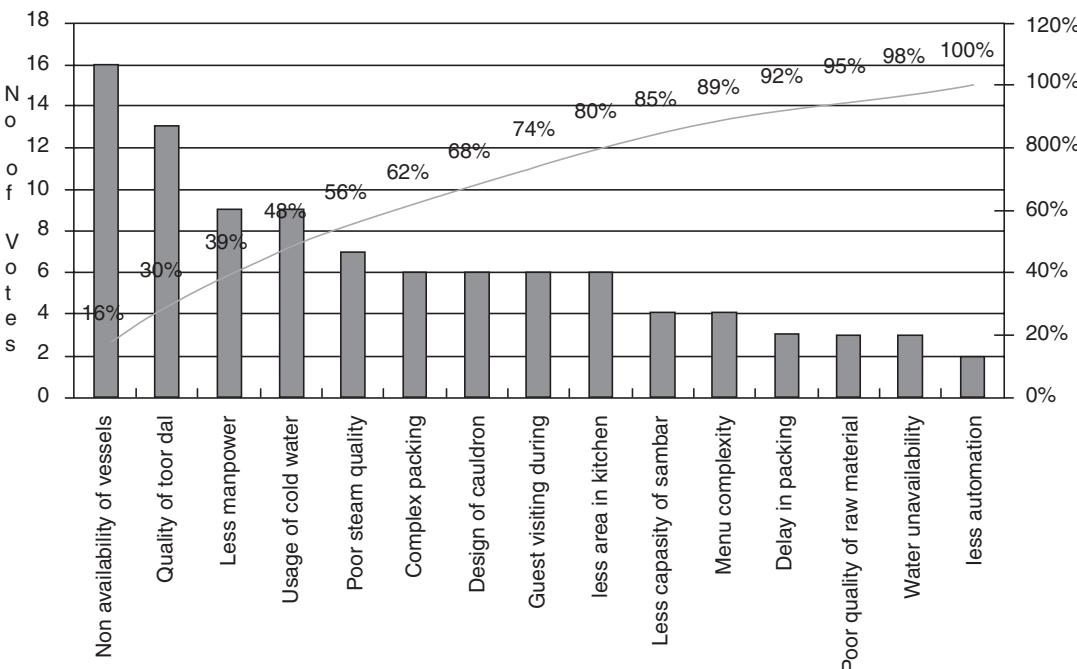
**EXHIBIT 8** Ishikawa's<sup>18</sup> cause-and-effect diagram for cooking cycle time reduction. Source: The Akshaya Patra Foundation.

**Note:** Note that 'Dal Boiling' and 'Seasoning Preparation' are parallel activities; these are carried out simultaneously with other cooking activities.

<sup>18</sup> The Fishbone Diagram is a tool for analyzing process dispersion. It is also referred to as the Ishikawa Diagram, because Kaoru Ishikawa developed it, and the Fishbone Diagram, because the complete diagram resembles a fish's skeleton. The diagram illustrates the main causes, which are divided into six Ms (Material, Method, Machine, Measurement, Man and Mother Nature), and sub-causes leading to an effect (symptom).

## Case Study

Continued...



**EXHIBIT 9** Pareto chart of potential causes as identified by the VK hill kitchen staff. Source: The Akshaya Patra Foundation.

### CASE QUESTIONS

- Using the information provided in the case, develop a high-level process map for the Akshaya Patra Foundation using Suppliers, Inputs, Process, Outputs, and Customers (SIPOC). What insights can you draw from the high-level process map?
- Given that VK Hill kitchen needs to serve 87,045 students with 7 tonnes of cooked rice and 16,800 liters of Sambar per day, what is the total time required for cooking Sambar and rice?
- If the food has to be delivered by 12.00 noon at the farthest school from the Vasanthapura kitchen (which takes 4 hours to reach), what are the lower and upper specification limits for completion of the first batch of Sambar? If the cooking starts at 5.00 am, calculate the corresponding Sigma Score (assume that the standard deviation for Sambar cooking is 10 minutes, while the average cooking time is as per **Exhibit 7**).
- For the upper and lower limits calculated in question 3 for Sambar cooking, calculate the optimal start time for Sambar cooking that will maximize the Sigma Score.
- If the food has to be delivered by 12.00 noon and the shortest delivery route takes about 2 hours, calculate the lower and upper specification limit for the second batch of Sambar. If the first batch of Sambar cooking time starts at the optimal start time identified in question 4, calculate the probability of not completing the second batch of Sambar cooking before upper specification limit for the second batch of Sambar cooking.
- Identify cost-effective means of reducing cooking cycle time, thereby increasing the Sigma Score.
- After improving the Sigma Score, what type of control mechanism needs to be adopted?
- What strategies TAPF should use to orient and motivate the unskilled workers of the kitchen on quality aspects?

## REFERENCES

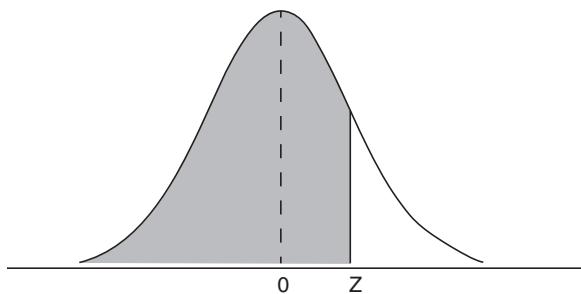
1. Anon (1988), "Motorola 1988 Annual Report", Available at [https://www.motorolasolutions.com/content/dam/msi/docs/enxw/static\\_files/1988\\_Motorola\\_Annual\\_Report.pdf](https://www.motorolasolutions.com/content/dam/msi/docs/enxw/static_files/1988_Motorola_Annual_Report.pdf). Accessed on 8 April 2017.
2. Anon (2004), "3M 2004 Annual Report", Available at [http://media.corporate-ir.net/media\\_files/NYS/MMM/reports/2004ar.pdf](http://media.corporate-ir.net/media_files/NYS/MMM/reports/2004ar.pdf), Accessed on 8 April 2017.
3. Anon (2006), "GSLV Failure due to Faulty Regulator", The Hindu Business Line, September 6 2016, available at <http://www.thehindubusinessline.com/todays-paper/tp-economy/gslv-failure-due-to-faulty-regulator-report/article1745318.ece> accessed on 18 May 2017.
4. Antony J and Banuelas R (2002), "Key Ingredients for the Effective Implementation of Six Sigma Program", *Measuring Business Excellence*, 6(2), 20–27.
5. Bowen D B and Headley D E (2002), *Airline Quality Rating 2002*, Available at URL <http://webs.wichita.edu/?u=aqr>.
6. Buss P and Ivey N (2001), "Dow Chemical Design for Six Sigma Rail Delivery Project", *Proceedings of the Winter Simulation Conference*, 1248–1251.
7. Chakravarty S N (1998), "Fast Food", *Forbes*, 8 October 1998, Available at <https://www.forbes.com/global/1998/0810/0109078a.html>. Accessed on 8 April 2017.
8. Duhigg C (2012), "The Power of Habit – What We Do and How to Change it", William Heinemann, London.
9. De Fao J and Bar-El Z (2002), "Creating Strategic Change More Efficiently with a New Design for Six Sigma Process", *Journal of Change Management*, 3(1), 60–80.
10. Kumar UD, Crocker J, Chitra T, and Saranga H (2006), "Reliability and Six Sigma", Springer, New York.
11. Kumar UD, Nowicki D, Ramirez-Marquez J E, and Verma D (2008), "On the Optimal Selection of Process Alternatives in a Six Sigma Implementation", *International Journal of Production Economics*, 111, 456–467.
12. Edgeman R L, Bigio D and Ferleman T (2005), "Six Sigma and Business Excellence: Strategic and Tactical Examination of IT Service Level Management at the Office of Chief of Technology Officer at Washington DC", *Quality and Reliability Engineering International*, 21, 257–273.
13. Hahn G J, Hill W J, Hoerl R W, and Zinkgraf S A (1999), "The Impact of Six Sigma Improvement – A Glimpse into Future of Statistics", *The American Statistician*, 53(3), 208–215.
14. Hendricks C A and Kelbaugh R L (1998), "Implementing Six Sigma at GE", *The Journal of Quality and Participation*, 21(4), 43–48.
15. Ishikawa K (1968), "Guide to Quality Control", Asian Productivity Organization, Tokyo.
16. Knowles G, Johnson M, and Warwood S (2004), "Medicated for Sweet Variability – A Six Sigma Application at a UK Food Manufacturer", *The TQM Journal*, 16(4), 284–292.
17. Kwak Y H and Anbari F T (2006), "Benefits, Obstacles and Future of Six Sigma", *Technovation: The International Journal of Technological Innovation, Entrepreneurship and Technology Management*, 26(5–6), 708–715.
18. Lanyon S (2003), "At Raytheon Six Sigma works too, to Improve HR Management Process", *Journal of Organizational Excellence*, 22(4), 29–42.
19. Lee K and Choi B (2006), "Six Sigma Management Activities and their Influence on Corporate Competitiveness", *Total Quality Management and Business Excellence*, 31(4), 833–863.
20. Liu E W (2006), "Clinical Research the Six Sigma Way", *The Journal of Association for Laboratory Automation*, 11(1), 42–49.
21. McClusky R (2002), "The Rise, Fall and Revival of Six Sigma", *Measuring Business Excellence*, 4(2), 6–17.
22. Mukhopadhyay A R and Ray S (2006), "Reduction of yarn packing defects using Six Sigma Methods", *Quality Engineering*, 18(2), 189–206.
23. Moore K (2011), "The Best Way to Innovation? An important Lesson from India", Forbes, May 24 2011. Available at <https://www.forbes.com/sites/karlmoore/2011/05/24/the-best-way-to-innovation-an-important-lesson-from-india/#2f12ae412861>, Accessed on 8 April 2017.
24. Motwani J, Kumar A, and Antony J (2004), "A Business Change Framework for Examining the Implementation of Six Sigma: A Case Study of Dow Chemicals", *TQM Magazine*, 16(4), 273–283.
25. Pyzdek T (2015), "The Six Sigma Handbook", 4<sup>th</sup> Edition, McGraw Hill, New York.
26. Subramanian T S (2006), "GSLV Crashes into Bay of Bengal", *The Hindu*, July 11, 2006.

27. Taguchi G (1986), “*Introduction to Quality Engineering – Designing Quality into Products and Processes*”, Asian Productivity Organization, Tokyo.
28. Taguchi G and Clausing D (1990), “Robust Quality”, *Harvard Business Review*, 65–75, January–February 1990.
29. Tomke S and Sinha M (2013), “The Dabbawala System: On-Time Delivery, Every Time”, *Harvard Business School Case* (Number 9-610-059).
30. Wyper B and Harrison A (2010), “Deployment of Six Sigma methodology in Human Resource Function”, *Total Quality Management*, 11(4 & 5), 720–727.
31. Wiener M (2004), “Six Sigma”, *Communication World*, 21(1), 26–29.
32. Zimmerman J P and Weiss J (2005), “Six Sigma’s Seven Deadly Sins”, *Quality*, 44(1), 62–66.

# Appendix

## Statistical Tables

Z - Table : Area Under the Standard Normal Curve

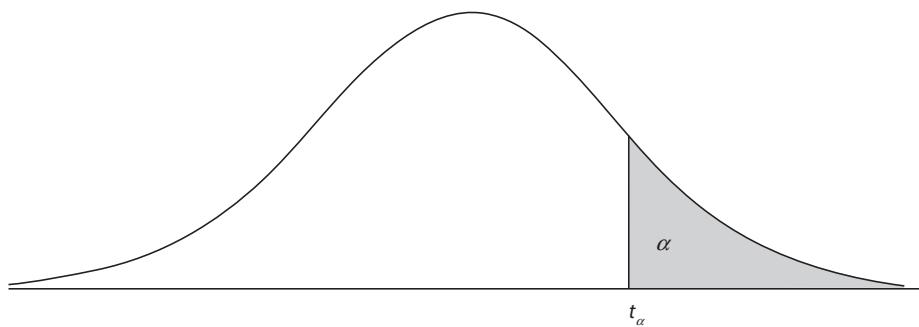


Number in the table represents :  $P(Z \leq z)$

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545

(Continued)

Number in the table represents :  $P(Z \leq z)$ —Continued

**t-Distribution Critical Values**Area to the right of:  $\alpha$ 

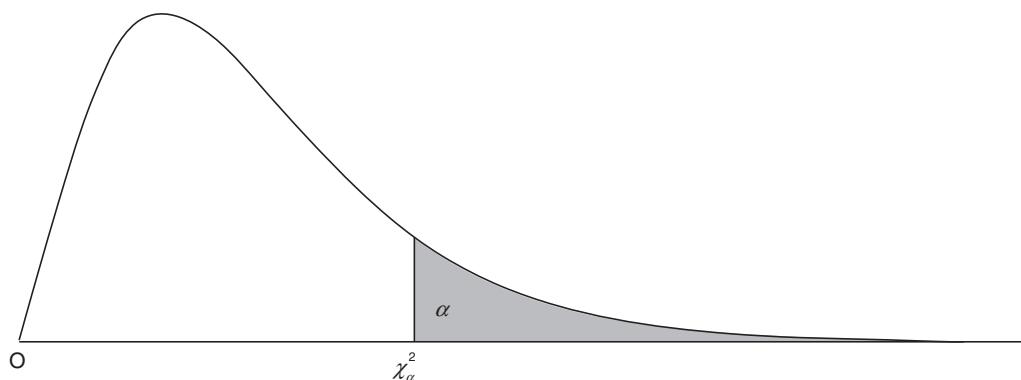
Degrees of freedom	$t_{0.100}$	$t_{0.050}$	$t_{0.025}$	$t_{0.01}$	$t_{0.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797

(Continued)

Area to the right of:  $\alpha$ —Continued

Degrees of freedom	$t_{0.100}$	$t_{0.050}$	$t_{0.025}$	$t_{0.01}$	$t_{0.005}$
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
50	1.299	1.676	2.009	2.403	2.678
60	1.296	1.671	2.000	2.390	2.660
100	1.290	1.660	1.984	2.364	2.626
120	1.289	1.658	1.980	2.358	2.617
$\infty$	1.282	1.645	1.960	2.326	2.576

## Critical Values of the Chi-Square Distribution



Degrees of freedom	$\chi^2_{0.995}$	$\chi^2_{0.990}$	$\chi^2_{0.975}$	$\chi^2_{0.950}$	$\chi^2_{0.900}$
1	0.0000393	0.0001571	0.0009821	0.0039321	0.0157908
2	0.0100251	0.0201007	0.0506356	0.1025866	0.2107210
3	0.0717218	0.1148318	0.2157953	0.3518463	0.5843744
4	0.2069891	0.2971095	0.4844186	0.7107230	1.0636232
5	0.4117419	0.5542981	0.8312116	1.1454762	1.6103080
6	0.6757268	0.8720903	1.2373442	1.6353829	2.2041307
7	0.9892557	1.2390423	1.6898692	2.1673499	2.8331069
8	1.3444131	1.6464974	2.1797307	2.7326368	3.4895391
9	1.7349329	2.0879007	2.7003895	3.3251128	4.1681590
10	2.1558565	2.5582122	3.2469728	3.9402991	4.8651821
11	2.6032219	3.0534841	3.8157483	4.5748131	5.5777848
12	3.0738236	3.5705690	4.4037885	5.2260295	6.3037961
13	3.5650346	4.1069155	5.0087505	5.8918643	7.0415046
14	4.0746750	4.6604251	5.6287261	6.5706314	7.7895336
15	4.6009156	5.2293489	6.2621378	7.2609439	8.5467562
16	5.1422054	5.8122125	6.9076644	7.9616456	9.3122364
17	5.6972171	6.4077598	7.5641864	8.6717602	10.0851863
18	6.2648047	7.0149109	8.2307462	9.3904551	10.8649361
19	6.8439714	7.6327296	8.9065165	10.1170131	11.6509100

(Continued)

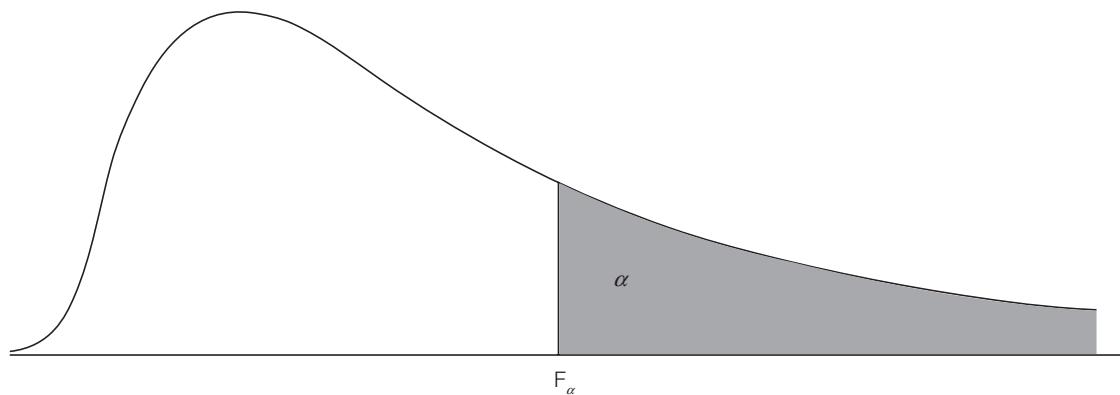
Degrees of freedom	$\chi^2_{0.995}$	$\chi^2_{0.990}$	$\chi^2_{0.975}$	$\chi^2_{0.950}$	$\chi^2_{0.900}$
20	7.4338443	8.2603983	9.5907774	10.8508114	12.4426092
21	8.0336534	8.8971979	10.2828978	11.5913052	13.2395980
22	8.6427164	9.5424923	10.9823207	12.3380146	14.0414932
23	9.2604248	10.1957156	11.6885519	13.0905142	14.8479558
24	9.8862335	10.8563615	12.4011502	13.8484250	15.6586841
25	10.5196521	11.5239754	13.1197200	14.6114076	16.4734080
26	11.1602374	12.1981469	13.8439050	15.3791566	17.2918850
27	11.8075874	12.8785044	14.5733827	16.1513958	18.1138960
28	12.4613359	13.5647098	15.3078606	16.9278750	18.9392424
29	13.1211489	14.2564546	16.0470717	17.7083662	19.7677436
30	13.7867199	14.9534565	16.7907723	18.4926610	20.5992346
40	20.7065353	22.1642613	24.4330392	26.5093032	29.0505229
50	27.9907489	29.7066827	32.3573637	34.7642517	37.6886484
60	35.5344911	37.4848515	40.4817480	43.1879585	46.4588883
70	43.2751795	45.4417173	48.7575648	51.7392780	55.3289396
80	51.1719319	53.5400773	57.1531729	60.3914784	64.2778445
90	59.1963042	61.7540790	65.6466176	69.1260304	73.2910905
100	67.3275633	70.0648949	74.2219275	77.9294652	82.3581358

### Critical Values of the Chi-Square Distribution—continued

Degrees of freedom	$\chi^2_{0.100}$	$\chi^2_{0.050}$	$\chi^2_{0.025}$	$\chi^2_{0.010}$	$\chi^2_{0.005}$
1	2.70554	3.84146	5.02389	6.63490	7.87944
2	4.60517	5.99146	7.37776	9.21034	10.59663
3	6.25139	7.81473	9.34840	11.34487	12.83816
4	7.77944	9.48773	11.14329	13.27670	14.86026
5	9.23636	11.07050	12.83250	15.08627	16.74960
6	10.64464	12.59159	14.44938	16.81189	18.54758
7	12.01704	14.06714	16.01276	18.47531	20.27774
8	13.36157	15.50731	17.53455	20.09024	21.95495

(Continued)

Degrees of freedom	$\chi^2_{0.100}$	$\chi^2_{0.050}$	$\chi^2_{0.025}$	$\chi^2_{0.010}$	$\chi^2_{0.005}$
9	14.68366	16.91898	19.02277	21.66599	23.58935
10	15.98718	18.30704	20.48318	23.20925	25.18818
11	17.27501	19.67514	21.92005	24.72497	26.75685
12	18.54935	21.02607	23.33666	26.21697	28.29952
13	19.81193	22.36203	24.73560	27.68825	29.81947
14	21.06414	23.68479	26.11895	29.14124	31.31935
15	22.30713	24.99579	27.48839	30.57791	32.80132
16	23.54183	26.29623	28.84535	31.99993	34.26719
17	24.76904	27.58711	30.19101	33.40866	35.71847
18	25.98942	28.86930	31.52638	34.80531	37.15645
19	27.20357	30.14353	32.85233	36.19087	38.58226
20	28.41198	31.41043	34.16961	37.56623	39.99685
21	29.61509	32.67057	35.47888	38.93217	41.40106
22	30.81328	33.92444	36.78071	40.28936	42.79565
23	32.00690	35.17246	38.07563	41.63840	44.18128
24	33.19624	36.41503	39.36408	42.97982	45.55851
25	34.38159	37.65248	40.64647	44.31410	46.92789
26	35.56317	38.88514	41.92317	45.64168	48.28988
27	36.74122	40.11327	43.19451	46.96294	49.64492
28	37.91592	41.33714	44.46079	48.27824	50.99338
29	39.08747	42.55697	45.72229	49.58788	52.33562
30	40.25602	43.77297	46.97924	50.89218	53.67196
40	51.80506	55.75848	59.34171	63.69074	66.76596
50	63.16712	67.50481	71.42020	76.15389	79.48998
60	74.39701	79.08194	83.29767	88.37942	91.95170
70	85.52704	90.53123	95.02318	100.42518	104.21490
80	96.57820	101.87947	106.62857	112.32879	116.32106
90	107.56501	113.14527	118.13589	124.11632	128.29894
100	118.49800	124.34211	129.56120	135.80672	140.16949

**Critical Values of the  $F$  Distribution ( $\alpha = 0.10$ )****Numerator Degrees of Freedom (DF1)**

	1	2	3	4	5	6	7	8	9
Denominator Degrees of Freedom (DF2)	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96

## Numerator Degrees of Freedom (DF1)—Continued

	1	2	3	4	5	6	7	8	9
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95
	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93
	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92
	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91
	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89
	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88
	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87
	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87
	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86
	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85
	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79
	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74
	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68
	$\infty$	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67

Critical Values of the F Distribution ( $\alpha = 0.10$ )—Continued

## Numerator Degrees of Freedom (DF1)

	10	12	15	20	24	30	40	60	120	$\infty$
Denominator Degrees of Freedom (DF2)	1	60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06
	2	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48
	3	5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14
	4	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.76
	5	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12
	6	2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74
	7	2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49
	8	2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32
	9	2.42	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18
	10	2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08
	11	2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00
	12	2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93
	13	2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88
	14	2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83

(Continued)

## Numerator Degrees of Freedom (DF1)—Continued

	10	12	15	20	24	30	40	60	120	$\infty$
15	2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.76
16	2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72
17	2.00	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.69
18	1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66
19	1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63
20	1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61
21	1.92	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59
22	1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.57
23	1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.55
24	1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.53
25	1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.52
26	1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	1.50
27	1.85	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.49
28	1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48
29	1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.47
30	1.82	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.46
40	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.38
60	1.71	1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.29
120	1.65	1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.26	1.19
$\infty$	1.60	1.55	1.49	1.42	1.38	1.34	1.30	1.24	1.17	1.00

Critical Values of the F Distribution ( $\alpha = 0.05$ )

## Numerator Degrees of Freedom (DF1)

	1	2	3	4	5	6	7	8	9	
Denominator Degrees of Freedom (DF2)	1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	

**Numerator Degrees of Freedom (DF1)—Continued**

	1	2	3	4	5	6	7	8	9
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96
$\infty$	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88

**Critical Values of the F Distribution ( $\alpha = 0.05$ )—Continued****Numerator Degrees of Freedom (DF1)**

	10	12	15	20	24	30	40	60	120	$\infty$
1	241.88	243.91	245.95	248.01	249.05	250.10	251.14	252.20	253.25	254.30
2	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50

(Continued)

## Numerator Degrees of Freedom (DF1)—Continued

	10	12	15	20	24	30	40	60	120	$\infty$
Denominator Degrees of Freedom (DF2)	3	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55
	4	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66
	5	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40
	6	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70
	7	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27
	8	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97
	9	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75
	10	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58
	11	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45
	12	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34
	13	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25
	14	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18
	15	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11
	16	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06
	17	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01
	18	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97
	19	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93
	20	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90
	21	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87
	22	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84
	23	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81
	24	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79
	25	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77
	26	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75
	27	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73
	28	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71
	29	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70
	30	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68
	40	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58
	60	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47
	120	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35
	$\infty$	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22

### Critical Values of the F Distribution ( $\alpha = 0.025$ )

Numerator Degrees of Freedom (DF1)

	1	2	3	4	5	6	7	8	9	
Denominator Degrees of Freedom (DF2)	1	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	
$\infty$	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	

**Critical Values of the F Distribution ( $\alpha = 0.025$ )—Continued****Numerator Degrees of Freedom (DF1)**

	10	12	15	20	24	30	40	60	120	$\infty$	
Denominator Degrees of Freedom (DF2)	1	968.63	976.71	984.87	993.10	997.25	1001.41	1005.60	1009.80	1014.02	1018.00
2	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50	
3	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90	
4	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26	
5	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02	
6	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85	
7	4.76	4.67	4.57	4.47	4.41	4.36	4.31	4.25	4.20	4.14	
8	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67	
9	3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33	
10	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08	
11	3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88	
12	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72	
13	3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60	
14	3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49	
15	3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40	
16	2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32	
17	2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25	
18	2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19	
19	2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13	
20	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09	
21	2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04	
22	2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00	
23	2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97	
24	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94	
25	2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91	
26	2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88	
27	2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85	
28	2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83	
29	2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81	
30	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79	
40	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64	
60	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48	
120	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31	
$\infty$	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00	

### Critical Values of the F Distribution ( $\alpha = 0.01$ )

Numerator Degrees of Freedom (DF1)

	1	2	3	4	5	6	7	8	9	
Denominator Degrees of Freedom (DF2)	1	4052	5000	5403	5625	5764	5859	5928	5981	6022
	2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
	3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
	4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
	5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
	6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
	7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
	8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
	9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
	10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
	11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
	12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
	13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
	14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
	15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
	16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
	17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68
	18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
	19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
	20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
	21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
	22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
	23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
	24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
	25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
	26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
	27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15
	28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
	29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09
	30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
	40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
	60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
	120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56
	$\infty$	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41

**Critical Values of the F Distribution ( $\alpha = 0.01$ )—Continued**

Numerator Degrees of Freedom (DF1)

	10	12	15	20	24	30	40	60	120	$\infty$	
Denominator Degrees of Freedom (DF2)	1	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
	2	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
	3	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
	4	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
	5	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
	6	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
	7	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
	8	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
	9	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
	10	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
	11	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
	12	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
	13	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
	14	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
	15	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
	16	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
	17	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
	18	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
	19	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
	20	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
	21	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
	22	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
	23	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
	24	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
	25	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
	26	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
	27	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
	28	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
	29	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
	30	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
	40	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
	60	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
	120	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
	$\infty$	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

**Bounds for Critical Values of the Durbin-Watson Statistic ( $\alpha = 0.025$ )**

	$k=1$		$k=2$		$k=3$		$k=4$		$k=5$	
$n$	$dl$	$du$								
15	1.08	1.36	0.95	1.54	0.81	1.75	0.69	1.98	0.56	2.22
16	1.11	1.37	0.98	1.54	0.86	1.73	0.73	1.94	0.62	2.16
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.66	2.10
18	1.16	1.39	1.05	1.54	0.93	1.70	0.82	1.87	0.71	2.06
19	1.18	1.40	1.07	1.54	0.97	1.69	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.89	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.65	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.87
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.75	1.00	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.14	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.53	1.35	1.59	1.30	1.65	1.24	1.72	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.20	1.79
39	1.44	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.37	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.53	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.52	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77

$k=1$		$k=2$		$k=3$		$k=4$		$k=5$		
$n$	$dl$	$du$	$dl$	$du$	$dl$	$du$	$dl$	$du$	$dl$	$du$
85	1.62	1.67	1.60	1.70	1.58	1.72	1.55	1.75	1.53	1.77
90	1.64	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.65	1.69	1.62	1.71	1.60	1.73	1.58	1.76	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

### Bounds for Critical Values of the Durbin-Watson Statistic ( $\alpha = 0.01$ )

$k=1$		$k=2$		$k=3$		$k=4$		$k=5$		
$n$	$dl$	$du$	$dl$	$du$	$dl$	$du$	$dl$	$du$	$dl$	$du$
15	0.81	1.07	0.70	1.25	0.59	1.47	0.49	1.71	0.39	1.97
16	0.84	1.09	0.74	1.25	0.63	1.45	0.53	1.66	0.44	1.90
17	0.87	1.10	0.77	1.26	0.67	1.43	0.57	1.63	0.48	1.85
18	0.90	1.12	0.81	1.26	0.71	1.42	0.61	1.60	0.52	1.80
19	0.93	1.13	0.84	1.26	0.74	1.42	0.65	1.58	0.56	1.77
20	0.95	1.15	0.86	1.27	0.77	1.41	0.68	1.57	0.60	1.74
21	0.98	1.16	0.89	1.28	0.80	1.41	0.72	1.55	0.63	1.71
22	1.00	1.17	0.91	1.28	0.83	1.41	0.75	1.54	0.67	1.69
23	1.02	1.19	0.94	1.29	0.86	1.41	0.78	1.54	0.70	1.67
24	1.04	1.20	0.96	1.30	0.88	1.41	0.81	1.53	0.73	1.66
25	1.05	1.21	0.98	1.31	0.91	1.41	0.83	1.52	0.76	1.65
26	1.07	1.22	1.00	1.31	0.93	1.41	0.86	1.52	0.78	1.64
27	1.09	1.23	1.02	1.32	0.95	1.41	0.88	1.51	0.81	1.63
28	1.10	1.24	1.04	1.33	0.97	1.41	0.90	1.51	0.83	1.62
29	1.12	1.25	1.05	1.33	0.99	1.42	0.92	1.51	0.86	1.61
30	1.13	1.26	1.07	1.34	1.01	1.42	0.94	1.51	0.88	1.61
31	1.15	1.27	1.09	1.35	1.02	1.43	0.96	1.51	0.90	1.60
32	1.16	1.28	1.10	1.35	1.04	1.43	0.98	1.51	0.92	1.60
33	1.17	1.29	1.11	1.36	1.06	1.43	1.00	1.51	0.94	1.59
34	1.18	1.30	1.13	1.36	1.07	1.44	1.01	1.51	0.95	1.59
35	1.19	1.31	1.14	1.37	1.09	1.44	1.03	1.51	0.97	1.59
36	1.21	1.32	1.15	1.38	1.10	1.44	1.04	1.51	0.99	1.59
37	1.22	1.32	1.16	1.38	1.11	1.45	1.06	1.51	1.00	1.59

(Continued)

	$k=1$		$k=2$		$k=3$		$k=4$		$k=5$	
$n$	$dl$	$du$								
38	1.23	1.33	1.18	1.39	1.12	1.45	1.07	1.52	1.02	1.58
39	1.24	1.34	1.19	1.39	1.14	1.45	1.09	1.52	1.03	1.58
40	1.25	1.34	1.20	1.40	1.15	1.46	1.10	1.52	1.05	1.58
45	1.29	1.38	1.25	1.42	1.20	1.47	1.16	1.53	1.11	1.58
50	1.32	1.40	1.29	1.45	1.25	1.49	1.21	1.54	1.16	1.59
55	1.36	1.43	1.32	1.47	1.28	1.51	1.25	1.55	1.21	1.59
60	1.38	1.45	1.35	1.48	1.32	1.52	1.28	1.56	1.25	1.60
65	1.41	1.47	1.38	1.50	1.35	1.53	1.31	1.57	1.28	1.60
70	1.43	1.49	1.40	1.51	1.37	1.55	1.34	1.58	1.31	1.61
75	1.45	1.50	1.42	1.53	1.40	1.56	1.37	1.59	1.34	1.62
80	1.47	1.52	1.44	1.54	1.42	1.57	1.39	1.60	1.36	1.62
85	1.48	1.53	1.46	1.55	1.43	1.58	1.41	1.60	1.39	1.63
90	1.50	1.54	1.47	1.56	1.45	1.59	1.43	1.61	1.41	1.64
95	1.51	1.55	1.49	1.57	1.47	1.60	1.45	1.62	1.43	1.64
100	1.52	1.56	1.50	1.58	1.48	1.60	1.46	1.63	1.44	1.65



# Bibliography

- Aczel, A. D. and Sounderpandian, J. (2008). *Complete Business Statistics* (7<sup>th</sup> edn.), Boston: McGraw Hill.
- Allen, D. M. (1971). Mean Square Error of Predictors as a Criterion for Selecting Variables, *Technometrics*, 13, 469–475.
- Anon (2010). Measurement Uncertainty – Principles and Methods, NASA Quality Assurance Handbook, NASA-HDBK-8739.19-3.
- Anscombe, F. J. and Tukey, J. W. (1963). The Examination of Analysis of Residuals, *Technometrics*, 5, 140–160.
- Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data* (3<sup>rd</sup> edn.), New York: John Wiley and Sons.
- Box, G. E. P. and Cox, D. R. (1964). An Analysis of Transformation, *Journal of Royal Statistical Society B*, 26, 211–243.
- Box, G. E. P. and Tidwell, P. W. (1962). Transformations of the Independent Variables, *Technometrics*, 4, 431–550.
- Buchanan, L. and O'Connell, A. (2006). A Brief History of Decision Making, *Harvard Business Review*, 2006, 32–41.
- Butcher, R. (1988). The Use of Postcode Address File as Sampling Frame, *Journal of Royal Statistical Society, Series D*, 37(1), 15–24.
- Cinlar, E. (1975). *Introduction to Stochastic Processes*, New York: Dover Publications.
- Cook, R. D. and Weisberg, S. (1983). Diagnostics for Heteroscedasticity in Regression, *Biometrika*, 70, 1–10.
- Draper, N. R. and Smith, H (1998). *Applied Regression Analysis* (3<sup>rd</sup> edn.), New York: John Wiley.
- Feller, W. (1957). *An Introduction to Probability Theory and its Applications – Volume I*, New Delhi: John Wiley & Sons.
- Feller, W. (1957). *An Introduction to Probability Theory and its Applications – Volume II*, New Delhi: John Wiley & Sons.
- Fisher, R. A. (1934). *Statistical Methods for Research Workers*, London: Oliver and Boyd.
- Hillier, F. S. and Lieberman, G. J. (2001). *Introduction to Operations Research*, New Delhi: Tata McGraw Hill.
- Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., and Tatham, R. L., (2012). *Multivariate Data Analysis* (7<sup>th</sup> edn.), New Delhi: Pearson.
- Hoaglin, D. C. and Welsch, R. (1978). The Hat Matrix in Regression and ANOVA, *The American Statistician*, 32, 17–22.
- Hocking, R. R. (1976). The Analysis and Selection of Variables in Linear Regression, *Biometrics*, 32, 1–49.
- Hosmer, D. W. and Lemeshow, S. (1980). Goodness of Fit Tests for the Multiple Logistic Regression Model, *Communications in Statistics A*, 9, 1043–69.
- Howard, R. A. (1971). *Dynamic Probabilistic Systems – Volume I Markov Models*, New York: John Wiley and Sons.
- Johnston, J. and DiNardo, J. (2007). *Econometric Methods* (4<sup>th</sup> edn.), Singapore: McGraw Hill International Editions.
- Kendall, M. G. (1938). A New Measure of Rank Correlation, *Biometrika*, 30(1), 81–93.
- Kleinbaum, D. G. and Klein (2011). *Logistic Regression – A Self Learning Text*, New Delhi: Springer.
- Lattin, J., Carroll, J. D. and Green, P. E. (2003). *Analyzing Multivariate Data*, New Delhi: Brooks/Cole.
- Liu, T. and Stone, C. C. (2006). Law and Statistical Disorder – Statistical Hypothesis Test Procedure and Criminal Trial Analogy, IDEAS URL <https://ideas.repec.org/p/bsu/wpaper/200601.html>.
- Marsland, S. (2015). *Machine Learning – An Algorithmic Perspective*, Florida: CRC Press.
- Pregibon, D. (1981). Logistic Regression Diagnostics, *Annals of Statistics*, 9, 705–724.
- Schapire, R. E. and Freund, Y. (2012). *Boosting – Foundations and Algorithms*, Cambridge, Massachusetts: MIT Press.
- Seber, G. A. F. and Lee, A. J. (2014). *Linear Regression Analysis* (2<sup>nd</sup> edn.), New Delhi: Wiley.
- Stahl, S. (2006). Evolution of Normal Distribution, *Mathematics Magazine*, 79(2), 96–113.
- Tate, R. F. and Klett (1959). Optimal Confidence Interval for the Variance of a Normal Distribution, *Journal of the American Statistical Association*, 54(287), 674–682.
- Thompson, B. (2006). Critique of p-Values, *International Statistical Review*, 74(1), 1–14.
- Winston, W. (2004). *Probability Models*, New Delhi: Cengage Learning.
- Winston, W. (2014). *Marketing Analytics – Data Driven Technique with Microsoft Excel*, New Delhi: Wiley.
- Witte, R. S. and Witte, J. S. (1997). *Statistics* (5<sup>th</sup> edn.), Florida: Harcourt Brace College Publishers.
- Wooldridge, J. M. (2006). *Introductory Econometrics – A Modern Approach*, New Delhi: Thomson South Western.



# Index

## A

accessible state, 594  
accuracy paradox in classification problem, 349  
adaptive boosting, 403  
adjusted  $R$ -square, 292  
agglomerative hierarchical cluster, 501  
Akshaya Patra Foundation (TAPF), 6, 666–682  
algebra of events, 60  
alternative hypothesis, 134, 136, 139  
Akaike Information Criteria, 466  
analysis of variance (ANOVA), 189–190, 245–246, 292–293  
    one-way, 192–194  
    setting up, 192–194  
    two-way, 199–200  
analytic hierarchy process and goal programming, 18  
analytics  
    reasons for use of, 4–7  
analytics capability building, 22–24  
    roadmap for, 24  
Anderson Goodman Test, 585  
Apache Hadoop, 19  
aperiodic state, 595  
area under the ROC curve (AUC), 352–354  
assignment problem, 526  
association rule learning, 64–65  
augmented Dickey–Fuller test, 465  
auto-correlation  
    defined, 296–297  
    Durbin–Watson test for, 297  
    of errors of lag 1, 297  
    of lag k, 297  
auto-correlation function (ACF), 452–453  
auto-regressive (AR) model, 451–452  
    partial auto-correlation function (PACF),  
        452–453  
auto-regressive integrated moving average (ARIMA) model,  
    450, 463–471  
    model building, 466–470  
auto-regressive moving average (ARMA), 450, 458  
axiomatic theory of probability, 61

## B

backward elimination method, 277, 303–304  
bank marketing data set, 365  
Bayes' theorem, 66  
    generalization of, 67  
    solving Monty hall problem using, 66–67  
Basic Variable, 531

Bayesian Information Criteria (BIC), 466  
Bernoulli trials, 72–74, 109  
Bessel's correction, 41  
Best Linear Unbiased Estimate (BLUE), 233, 278  
big data analytics, 18–19  
binary logistic regression, 338–339  
    estimation of parameters, 340  
    log-likelihood function of, 340  
    probability (likelihood) function of, 340  
binding constraint, 530  
binomial distribution, 73–75, 167  
    approximation using normal distribution, 75  
    cumulative distribution function of, 74–75  
    mean of, 75  
    probability mass function of, 74  
    variance of, 75  
black belt projects, 653  
blending problem, 526  
Bonferroni Correction, 393, 395–396  
bootstrap aggregating (known as bagging), 105, 404  
Box Plot, 51  
branch and bound algorithm, 554–555  
breadth first search (BFS), 558  
Breaking Barriers: Micro-mortgage analytics, 412–420  
business intelligence (BI), 32

## C

categorical variables, 231, 288–289  
cause-and-effect diagram (fish bone diagram or Ishikawa diagram), 662  
central limit theorem (CLT), 87, 108–109, 141, 161  
    alternative version of, 108  
    implications of, 108  
    for population proportion, 127  
    for proportion, 109  
Central Parking Services Private Limited, 178–185  
Chapman–Kolmogorov equation, 584  
Chebyshev's theorem, 43  
children nodes, 398  
chi-square automatic interaction detection (CHAID), 90,  
    392–398  
business rules, 396–398  
contingency table, 393, 394  
initial models, 392  
internal nodes, 393  
nature of dependent variables in, 393  
null and alternative hypotheses in, 394  
root node, 393–394

- statistic for Chi-square Test of Independence, 394  
 steps for developing, 393–395  
 terminal nodes, 393, 395, 397  
**chi-square distribution**, 90–92  
 confidence interval for, 129–130  
 cumulative distribution function of, 91  
 degree of freedom, 90  
 probability density function of, 90–91  
 properties of, 92  
**chi-square goodness of fit tests**, 166–167  
 choice of number of intervals in, 167  
**chi-square statistic value**, 173–174  
**chi-square test of independence**, 172–174  
**classification and regression trees (CART)**, 392  
**classification cut-off probability**, 347–348, 349  
 based on penalty cost, 364–365  
**classification problems**, 337–338  
 accuracy paradox in, 349  
 decision trees for solving, 391  
 objectives of, 338–339  
 tradeoff between sensitivity and specificity in, 350  
**classification table in a logistic regression model**, 347–348  
**clickstream data**, 33  
**clustering**, 489–490  
 algorithms, 497–501  
 cluster validation, 498  
 deciding distance/similarity measures, 497–498  
 distance and dissimilarity measures used, 490–496  
 hierarchical, 501–504  
 K-means, 498–501  
 quality and optimal number of clusters, 496–497  
 variable selection, 497  
**cluster sampling**, 105  
**Cochran's theorem**, 194  
**coefficient of determination**, 241  
**co-efficient of multiple determination**  
 ( $R^2$  or  $R^2$ ), 291–292  
**Cohen's D**, 162–163  
**combinatorial optimization**, 18  
**communicating states**, 594  
**complementary slackness theorem**, 535, 545  
**concordant pair**, 350–351  
**conditional probability**, 137  
 of event, 63–64  
**confidence interval**  
 case study, 254–262  
 for expected value of  $Y$  for a given  $X$ , 252  
 for population mean, 124–126, 128–129  
 for population proportion, 127  
 for population variance, 129–130  
 for regression coefficients  $\beta_0$  and  $\beta_1$ , 251–252  
**confusion matrix**, 348  
**consistency**, 112  
**contingency table**, 172  
**continuous distributions**, parameters of, 81–82  
**continuous random variables**, 70  
 cumulative distribution function, 72–73  
 probability density function, 72–73  
 variance of, 73  
**constraint**, 528  
**customer lifetime value**, 577, 601  
**convenience sampling**, 106  
**Cook's distance**, 250, 298  
**correlation coefficient**, 208  
**correlogram**, 452  
**cosine similarity**, 494–495  
**cost-based cut-off probability**, 357–358  
**cost-based splitting criteria**, 402–403  
**cost of poor quality (CoPQ) measures**, 639–340  
**costs of decision making**, 5  
**Cox and Snell  $R^2$** , 347  
**credit rating**, application of logistic regression (LR) in, 359–362  
**credit score using LR**, 362–365  
**credit worthiness of a customer**, 362  
**critical value**, 110, 138  
**cross-sectional data**, 34  
**Croston's method for intermittent demand**, 441–442  
**cumulative distribution function**  
 binomial distribution, 74–75  
 Chi-square distribution, 91  
 discrete random variables, 70–71  
 F-distribution, 94  
 geometric distribution, 80  
 Poisson distribution, 77–78  
 uniform distribution, 82  
**cumulative television rating points (CTRP) of a television program**, 281–284  
**customer analytics at Flipkart.com**, 619–629  
**cutting stock problem**, 526
- D**
- data**  
 cross-sectional, 34  
 panel, 34  
 structured and unstructured, 32–33  
 time series, 34  
**data-driven decision making**, 2–4  
 business context, 8–9  
 challenges in, 25–27  
 data science, 9–10  
 framework for, 22  
 technology, 9  
 through business analytics, 7  
**data extraction, transformation, and loading (ETL)**, 2  
**data imputation techniques**, 231  
**data measurement scales**  
 interval scale, 35  
 nominal scale, 34  
 ordinal scale, 34–35  
 ratio scale, 35  
**data science**, 7–10, 27  
**data scientist**, 1  
**data visualization**  
 bar chart, 48  
 box plot (aka box and whisker plot), 51  
 Coxcomb chart, 50

- histogram, 45–48  
 pie chart, 49  
 scatter plot, 49–50  
 Treemap, 51–52  
 decile, 38  
 decision-making process using past data, 2–3  
 decision trees, 17  
     criteria for developing, 392  
     divide-and-conquer strategy, 391–392  
     internal nodes, 392  
     merging criteria, 392  
     root node, 392  
     for solving classification problems, 391  
     splitting criteria, 392  
     steps for generating, 392  
     stopping criteria, 392  
     terminal nodes (aka leaf nodes), 392  
     tree pruning, 392  
 defects per million opportunities (DPMO), 648–649  
 Define, Measure, Analyse, Improve, and Control (DMAIC), 653–655  
     armoured vehicles, case of, 655–661  
 degrees of freedom, 42, 245  
 DeMorgan's Laws, 60  
 depth first search (DFS), 558  
 descriptive analytics, 10–13, 32, 232  
     primary objective of, 32  
     techniques, 17–18  
 DFBeta, 251, 301–302  
 DFFit, 251, 299–300  
 Dickey–Fuller test, 464–465  
     augmented, 465  
 difference stationarity, 465  
 discordant pair, 350–351  
 discrete random variables, 69  
     cumulative distribution function of, 70–71  
     expected value (or mean) of, 71  
     probability mass function of, 70–71  
     standard deviation of, 71–72  
     variance of, 71–72  
 distance measures  
     Cook's distance, 298  
     critical values of various, 302  
     Euclidian distance, 298  
     leverage value (or hat value), 299  
     Mahalanobis distance, 298–299  
 divide-and-conquer strategy, 391–392  
 double exponential smoothing, 437–438  
 downward bias, 42  
 dummy variables, 231, 286, 288  
 Durbin–Watson statistic value, 444  
 Durbin–Watson test, 297  
 dynamic programming, 609–610
- E**  
 e-commerce companies, 2–3  
 ensemble method, 403–404  
 entropy, 401
- ergodic Markov chain, 596  
 ERP (enterprise resource planning) systems, 230  
 error function, 85  
 error matrix, 348  
 estimation of parameters  
     of binomial distribution parameter, 115–116  
     of exponential parameter, 117–118  
     using maximum likelihood estimation (MLE), 115  
     using method of moments, 114  
 E-tailers, 524  
 Euclidean distance, 298, 490–493  
     standardized, 493
- events, 58  
     algebra of, 60  
     conditional probability of, 63–64  
     independent, 63  
     joint probability of, 61–62  
     probability estimation of, 59
- evolutionary learning algorithms, 22
- explanatory variable, 227
- exponential distribution, 82–83  
     cumulative distribution of, 82  
     mean of, 82  
     memoryless property of, 83  
     probability density function of, 82  
     variance of, 82
- exponential smoothing, 435  
     single exponential smoothing (SES), 435–437  
     triple exponential smoothing (Holt–Winter model), 437–441
- F**  
 Facebook, 2, 20  
 Facebook Relationship Breakups, 10–11  
 factor effect model, 190  
 farm advisory systems, 3  
 F-distribution, 94–95  
     cumulative distribution function of, 94  
     degrees of freedom, 94  
     probability density function of, 94  
     properties of, 95  
 feasible region, 530  
 feature selection, 208  
 first-order Markov process, 582  
 first-order moment, 114  
 first passage time, 595  
 Fisher, R A, 192  
 Five whys (or 5 whys), 663  
 Florence Nightingale, 50  
 forecasting, 427  
     Croston's method for intermittent demand, 441–442  
     regression model for, 444–446  
     technique and accuracy, 430–431  
     Theil's coefficient, 471  
     time-series data with seasonal variation, 445–446
- forward LR (likelihood ratio), 358  
 forward selection method, 277, 303  
 Forward Selection Wald, 358–359  
 fraudulent transactions, 2–3

*F*-Score (*F*-Measure), 350  
*F*-statistic value, 246, 293  
*F*-test, 245–246, 292–293  
 full model, 190  
 functional form of regression model, 248–249  
 functional form of relationship, 232–233

**G**

Gain chart, 365–369  
 Galton, Francis, 228, 231  
 Gamma function, 90, 91  
 generalized linear models (GLM), 339  
 geometric distribution, 79–81  
   cumulative distribution function of, 80  
   mean of, 80  
   memoryless property, 80–81  
   probability density function of, 79  
   probability mass function of, 80  
   variance of, 80  
 Gini coefficient, 352–353  
 Gini impurity index, 398–401  
 goal programming, 558–559  
 Gower's similarity coefficient, 495–496  
 green belt projects, 653

**H**

hat matrix, 274  
 heteroscedasticity, 234, 273  
 hierarchical clustering, 490, 501–504  
 HiPPO algorithm (“highest paid person’s opinion” algorithm), 3  
 Holt’s method, 437–438  
 Holt–Winter model, 437–441  
 Homogeneous Poisson Process (HPP), 579  
 homoscedasticity, 234, 273  
   test of, 248  
 Hosmer–Lemeshow (H–L) test, 344, 346–347  
 HR Analytics at ScaleneWorks, 381–390  
 hypothesis test for regression co-efficient (*t*-test), 243–245  
 hypothesis testing  
   blackout babies, 134  
   comparing *Z*-test and *t*-test, 156–160  
   for correlation coefficient, 213–214  
   critical value, 138  
   decisions in, 141  
   description of hypothesis, 136  
   for difference in population proportion under large samples, 161  
   for equality of population variances, 163  
   for one-sample test for proportion, 149  
   one-sample *Z*-test, 141–143  
   one-tailed and two-tailed test, 138–140  
   paired *t*-test, 154  
   power of test and power function, 148–149  
   rejection region, 138  
   setting up, 135–138  
   significance value, 138  
   *t*-test, 151

two-sample *t*-test, 158  
 two-sample *t*-test with unequal variance, 159–160  
 two-sample *Z*-test, 156–157  
 type I and type II errors, 141

**I**

independent and identical distribution (IID), 108  
 independent events, 63  
 influence matrix, 274  
 influential observation, 298  
 integer linear programming (ILP), 553–558  
 integer programming model, 17  
 interaction variables, 276  
   in regression models, 289  
 internal nodes, 392, 393  
 Inter quartile distance (IQD), 40  
 interval estimate, 111  
   definition, 124  
 iso-profit line, 537–538

**J**

Jaccard similarity coefficient (JSC), 493–494  
 Jesus Christ, 101  
 John Snow, 11–12  
 John Snow’s Spot Map, 11–12  
 joint probability of events, 61–62

**K**

*k*-degrees of freedom, 166  
 key performance indicators (KPIs), 2, 32, 226, 271, 275  
 K-means, 498–501  
 Kolmogorov, Andrey, 61  
 kurtosis, 44

**L**

Larsen and Toubro - Spare Parts Forecasting, 477–487  
 LASSO regression, 296  
 left-tailed test, 140  
 leptokurtic distribution, 44  
 leverage value of an observation, 250–251  
 lift chart, 365–369  
 likelihood of observing evidence, 66  
 likelihood ratio test, 589  
 limiting probability, 596  
 linear programming, 17, 524–527  
   assumptions of, 531–532  
   dual, 540–542  
   model building, 527  
   sensitivity analysis in, 532–535  
   solving using graphical method, 535–539  
   terminologies, 530–531  
 Linen management at Apollo Hospital, 570–575  
 link function, 339  
 Literary Digest, 100, 102  
 Ljung–Box Test for Auto-Correlations, 470–471  
 location parameter, 82  
 location problem, 526  
 logistic and multinomial regression, 17

- logistic regression (LR), 337–338  
 accuracy in predicting positive classes, 349  
 accuracy of classifying negatives, 348  
 application in credit rating, 359–362  
 binary, 338–339  
 classification table in, 347–348  
 credit score using, 362–365  
 estimation of parameters in, 340  
 interpretation of parameters in, 342–343  
 model diagnostics, 343–347  
 overall accuracy of, 348  
 precision measures, 350  
 sensitivity, 349–350  
 specificity, 349–350  
 variable selection in, 358–359
- logit (logistic probability unit) function, 339  
 definition, 339  
 S-shaped curve, 339
- log likelihood function, 118
- Lorenz, Max O, 353
- Lorenz curve, 353
- Lower control limit (LCL), 646–647
- Lower specification limit (LSL), 646–647
- M**
- machine learning algorithms, 21–22
- Mahalanobis distance, 250, 298, 299
- Mallows's  $C_p$ , 277, 305–306
- Manhattan distance (city block distance), 493
- marginal probabilities, 63
- Markdown Optimization, 507–508
- Markov chain, 17, 583–585  
 with absorbing states, 596–597  
 expected duration to reach a state of, 599–600  
 in predictive analytics, 589–590  
 retention probability and customer lifetime value using, 601–602  
 stationary distribution in, 590–591
- Markov decision process (MDP), 578, 603–604
- Markov property, 83
- materials requirements planning (MRP), 427
- matrix representation of multiple regression, 272
- maxima of the log likelihood function, 119
- maximum likelihood estimation (MLE), 115, 123, 340  
 of normal distribution parameters, 118–119
- maximum value of Cox and Snell  $R^2$ , 347
- mean absolute error (MAE), 431
- mean absolute percentage error (MAPE), 431
- mean (or average) value, 36–37  
 binomial distribution, 75  
 exponential distribution, 82  
 geometric distribution, 80  
 uniform distribution, 82
- mean recurrence time, 595
- means model, 190
- mean square error (MSE), 250, 431
- mean time between failure (MTBF), 100
- measure of impurity, 398
- measures of central tendency, 35–38
- measures of variation, 40–43
- median (or mid) value, 37
- memoryless property of geometric distribution, 80–81
- merging criteria, 392
- mesokurtic distribution, 44
- method of moments, 113–114
- Minkowski distance, 493
- mixed integer programming, 553
- mode, 38
- model building, 208
- model building using simple linear regression, 228–234
- Monty Hall problem, 5–6
- moving average (MA) processes, 457–458
- moving average method, 432
- multi-collinearity, 208, 273  
 correlation between independent variables, 295  
 impact on MLR model, 295  
 LASSO regression for handling, 296  
 Principle Component Analysis (PCA) for handling, 296  
 remedies for handling, 296  
 Ridge regression for handling, 296  
 variance inflation factor (VIF) and, 295–296
- multi-criteria decision-making (MCDM) model, 18, 558–559
- multi-period (stage) models, 551
- multiple linear regression (MLR) model, 271  
 assumptions of, 272–273  
 backward elimination method, 277, 303–304  
 building a, 275–279  
 co-efficient of multiple determination ( $R$ -Square or  $R^2$ ), 291–292  
 data extraction and collection, 276  
 data quality, 276  
 deriving new variables, 276  
 estimating regression parameters, 278  
 forward selection method, 277, 303  
 framework for developing, 275  
 functional form, defining, 277  
 handling qualitative variables, 276  
 identifying proxy variables, 277  
 implementation of, 278–279  
 interaction variables, 276, 289  
 interpretation of, 282–284  
 missing data, 276  
 ordinary least squares estimation for, 272–275  
 part (semi-partial) correlation and, 279–281  
 performing diagnostics, 278  
 pre-processing of data, 276  
 with qualitative variables, 285–289  
 residual analysis, 294  
 statistical significance of individual variables in, 292  
 stepwise regression, 277, 304–305  
 training and validation sets, 277  
 transformation in, 306, 312–313  
 with two explanatory variables, 295–296  
 validating of, 278, 291, 293–294  
 variable selection in, 303–305  
 multiplicative time-series model, 429

**N**

- Nagelkerke  $R^2$ , 347
- Naïve forecasting model, 471
- negative outcome, 338
- non-basic variable, 531
- non-binding constraint, 531
- non-linear programming (NLP), 18
- non-linear regression, 227
- non-overlapping clusters, 490
- non-preemptive goal programming, 559
- non-probability sampling, 105–106
- non-stationary process into a stationary process,
  - transforming, 465
- normal distribution (Gaussian distribution), 85–88
  - confidence interval for, 128–129
  - likelihood function of, 118
  - properties of, 86–87
- null hypothesis, 134, 136, 139
- null-recurrent state, 595

**O**

- omnibus test (likelihood ratio test), 344–345
- one-sample test for proportion, 149
- one-sample Z-test, 141–143
- one-step transition probabilities of Markov chain, 584–585
- one-tailed test, 138–140
- one-way ANOVA, 192–194
- opinion-based decision making, 2
- optimal cut-off probability, 350
  - classification plot for selection of, 354–355
  - cost-based approach, 357–358
  - Youden's Index for, 356–357
- optimal policy, linear programming for finding, 608–609
- optimal smoothing constant in SES, 437
- ordinary least squares (OLS), 233, 278
  - estimation for multiple linear regression (MLR) model, 272–275
  - estimation of parameters, 234–236
  - estimation of regression parameters using, 340
- outcome variable, 227
- outlier analysis, 249–251
- outliers, diagnosing, 297–299
- Out-of-Bag (OOB) data, 405
- overlapping clusters, 490

**P**

- paired t-test, 154
- panel data, 34
- partial auto-correlation function (PACF), 452–453
- partial correlation, 279–280
- partial F-test, 293–294
- partial regression coefficients, 274, 282–284
- Pearson correlation coefficient, 208–214, 236, 280
  - properties of, 211
  - value of, 210
- Pearson product moment correlation coefficient, 209–210

- Pearson's moment coefficient of skewness, 43–44
- percentile, 38
- periodic state, 595–596
- Phi-coefficient, 218–219
- platykurtic distribution, 44
- point bi-serial correlation, 217
- point estimate, 111
- Poisson distribution, 77–78
  - cumulative distribution function of, 77–78
  - probability mass function of, 78
- Poisson process, 578–579
  - compound, 582
- policy iteration algorithm, 605–606
- population, 35
- population parameters, 101
  - estimation of, 111–113
- positive outcome, 338
- positive recurrent state, 595
- posterior probability, 66
- P-P plot (Probability- Probability plot), 246, 247
- precision measures, 350
- prediction interval for value of Y for a given X, 253
- predictive analytics, 13–15, 207, 232
  - techniques, 17–18
- predictor variable, 227
- preemptive goal programming, 559
- pregnancy prediction, 8
- prescriptive analytics, 15–16, 523–524
  - techniques, 17–18
- primal-dual relationships, 544–546
- principal component analysis (PCA), 296
- prior probability, 66
- probabilistic clusters, 490
- probability density function
  - Chi-square distribution, 90–91
  - continuous random variables, 72–73
  - exponential distribution, 82
  - F-distribution, 94
  - geometric distribution, 79
  - student's t-distribution, 93
  - uniform distribution, 82
- probability estimation of an event, 59
- probability mass function
  - binomial distribution, 74
  - discrete random variables, 70–71
  - geometric distribution, 80
  - Poisson distribution, 78
- probability sampling, 103–105
- probability theory
  - terminology, 58–60
- process capability, 633, 636, 643–646
- product mix problem, 526
- profit maximization, 5
- pseudo  $R^2$ , 347
- pure integer programming, 553
- p-value, 137, 296

**Q**

$Q$ -statistic, 471  
qualitative variables of multiple linear regression (MLR)  
model, 285–289  
quartile, 38

**R**

random experiment, 58  
Random Forest, 17, 404–405  
random sampling, 103–104  
random variables  
continuous, 70  
definition of, 68  
discrete, 69  
range of optimality, 539–540  
Rattle Package, 405  
receiver operating characteristic (ROC) curve, 352–353, 356, 405  
recurrent state, 594–595  
Reduced cost, 523, 531–532  
regression, 17  
regression coefficients of categorical variables, 288–289  
regression-Francis Galton's regression model, 228  
regression model for forecasting, 444–446  
regression parameters, estimate of, 233  
Regression Tree, 398  
regular matrix, 592  
reinforcement learning algorithms, 22  
rejection region, 138  
of a two-tailed test, 140  
relative frequency, 59  
remedial measures diagnostics, 233  
repeated trials, 141  
residual (error) analysis, 246–248  
response surface (hyperplane), 272  
response variable, 227  
ridge regression, 296  
right-tailed test, 139  
rolled throughput yield, 651  
root mean square error (RMSE), 278, 431  
root node, 392, 393–394, 398  
 $R$ -square or  $R^2$ , 444  
adjusted, 292  
co-efficient of multiple determination, 291–292  
Cox and Snell, 347  
Nagelkerke, 347  
pseudo, 347  
value, 240–243  
Rubin Causal Model, 226  
100% rule for simultaneous changes, 534–535

**S**

sample, 35  
size, estimation for mean of the population, 110–111  
space, 58  
statistic or statistic, 101  
variance ( $S^2$ ), 41  
sampling, 99–100

cluster, 105

convenience, 106  
distribution, 106–108  
non-probability, 105–106  
random, 103–104  
respondent bias in, 101  
steps, 102  
stratified, 104  
voluntary, 106  
sampling distribution  
of correlation coefficient, 214  
of Spearman correlation, 215

scale parameter, 81

seasonality index using method of averages, 440

second-order moment, 114

semi-partial correlation (or part correlation), 280–281

sensitivity, 349–350

set covering problem, 526

set theory relationships, 61

shadow price, 533

range of, 540

shape, measures of, 43–45

shape parameter, 82

Sigma Score, 640–641, 651–652

significance, 124

significance value, 138

sign of correlation coefficient, 211

Simon, Herbert, 2

simple linear regression, 225

interpretation of regression coefficients, 238–240

model building, 228–234

validation of, 240–249

simple linear regression (SLR), 226–227

simple moving average (SMA), 432

single exponential smoothing (SES), 435–436

optimal smoothing constant in, 437

single factor experimental studies, 190

Six Sigma, 18, 633–635

definition, 636

industrial applications of, 641–643

measures, 643–647

Mumbai Dabbawalas case, 633–635

origins of, 636–638

three-sigma process, 638–639

toolbox, 661–663

slack variable, 531

skewness, 43–44

smoothing constant, 435

social media analytics, 18

Spearman rank correlation, 215

specificity, 349–350

splitting criteria, 392

SPSS logistic regression, 366

spurious correlation, 213

spurious regression, 242–243

s-step transition probability, 584

standard deviation, 40–42

- point bi-serial correlation, 217  
 standardized beta, 285  
 standardized dependent variable, 285  
 standardized DFFIT (SDFFIT), 299–300  
 standardized independent variables, 285  
 standardized regression coefficients, 284–285  
 standard normal distribution, 137  
 standard normal variable, 87–88  
 state space, 578, 582  
 statistically significant, 288  
 of logistic regression model, 343  
 multi-collinearity and, 295  
 statistical significance of individual variables in *F*-test, 292  
 stepwise regression, 277, 304–305  
 stochastic process models, 578  
 stock keeping units (SKUs), 3, 59, 428  
 stopping criteria, 392  
 stratified sampling, 104  
 stratum, 104  
 strong law of duality, 545  
 structured data, 32–33  
 student's *t*-distribution, 92–94  
 cumulative distribution function of, 93  
 degrees of freedom, 92  
 mean, 92  
 probability density function of, 93  
 properties of, 94  
 standard deviation, 92  
 sum of squares errors (SSE), 235, 273, 291–292, 398  
 sum of squares of between (SSB) group variation, 193–194  
 sum of squares of total variation (SST), 193  
 sum of squares of within the group variation, 194  
 supervised learning algorithms, 22  
 Suppliers, Inputs, Process, Output, and Customers (SIPOC),  
   662–663  
 surplus variable, 531
- T**  
 Target's pregnancy test, 8–9  
*t*-distribution  
   with (*n*-1) degrees of freedom, 128  
   with (*n*-2) degrees of freedom, 214  
 terminal nodes (aka leaf nodes), 392, 393, 395, 397, 398  
 test statistic, 137–138  
 Theil's coefficient, 471  
 theory of bounded rationality, 2  
 theory of firm, 4–5  
 threshold value for VIF, 296  
 time series data, 34  
 time-series data, 428–430  
   cyclical component, 429  
   irregular component, 429  
   seasonal component, 429  
   trend component, 429  
 tolerance, 296  
 training data set, 231  
 transformation in MLR, 306, 312–313
- Tukey's Bulging Rule, 313  
 transient state, 595  
 transportation problem, 526  
 travelling salesman problem (TSP), 6–7  
 tree pruning, 392  
 trend stationarity, 465  
 triple exponential smoothing (Holt–Winter model), 437–441  
*t*-statistic, 151, 296  
 auto-correlation and, 314  
 multi-collinearity and, 295  
 for null hypothesis, 214  
 for the variables *CTRP*, 292  
 variance inflation factor (VIF) and, 296–297  
 Tukey's Bulging rule for transformations, 313  
 two-sample *t*-tests, 191, 195  
 two-tailed test, 138–140  
 type I and type II errors, 141
- U**  
 unbiased estimate of a population parameter, 111  
 uniform distribution, 82  
 cumulative distribution function of, 82  
 mean of, 82  
 probability density function of, 82  
 variance of, 82  
 unstructured data, 32–33  
 unsupervised learning algorithms, 22  
 upper control limit, 646–647  
*U*-statistic, 471
- V**  
 validation data set, 231, 233  
 validation of multiple linear regression (MLR) model,  
   278, 291, 293–294  
 value iteration algorithm, 609–610  
 variable selection, 208  
 variance for population, 40–42  
 variance inflation factor (VIF), 278, 295–296  
   threshold value for, 296  
 vehicle routing problem (VRP), 6  
 Venn diagram, 59  
   partial correlation, 280  
   semi-partial correlation (or part correlation), 280–281  
 voluntary sampling, 106
- W**  
 Wald's test, 344, 346  
 weak law of duality, 544  
 web and social media analytics, 19–21
- Y**  
 yield (*Y*), 650–651  
 Youden's Index, 356–357, 363
- Z**  
 zero-one (or binary) programming, 554  
 Z-score, 250  
 Z-test, 141–143

# Business Analytics

The Science of  
Data-Driven Decision Making

U Dinesh Kumar

## About the Book

Written with the aim of becoming the primary resource for students of business analytics, this book provides a holistic perspective of analytics with theoretical foundations and applications of the theory using examples across several industries. The content of the book starts with the foundations of data science and includes descriptive, predictive and prescriptive analytics topics which are discussed using examples from several industries as well as nine analytics case studies distributed by Harvard Business Publishing and used by several institutes across the world. The book is enriched with 10 years of teaching experience of the author at various programs in the Indian Institute of Management Bangalore and several training programs and consulting projects carried out by him.

## Key Features

- Equal importance to theory and practice with examples across industries.
- Case studies providing deeper understanding of analytics techniques and deployment of analytics driven solutions.
- Coverage of topics such as basic probability concepts, probability distributions, hypothesis testing, multiple linear regression, logistic regression, decision trees, forecasting, clustering, prescriptive analytics, stochastic models and Six Sigma.
- Data sets have been hosted in Wiley's Web page: <https://www.wileyindia.com/business-analytics-the-science-of-data-driven-decision-making.html>
- Discussion of analytics applications in industries such as banking and finance, e-commerce, healthcare, manufacturing, retail and services.

*The strongest point of this book is that it blends both theory and applications very well.*

—Professor M K Tiwari, IIT Kharagpur

*This book is needed since many books in the market are either purely theoretical or practitioner oriented.*

—Dr M Mathirajan, Indian Institute of Science Bangalore

*There are not many books with the level of comprehensiveness that Prof Dinesh's book provides.*

—Kiran R, Director Data Sciences and Advanced Analytics, VMWare

follow us on

 [facebook.com/wileyindia](http://facebook.com/wileyindia)

 [twitter.com/wileyindiaapl](http://twitter.com/wileyindiaapl)

 [linkedin.com/in/wileyindia](http://linkedin.com/in/wileyindia)

 [google.com/+wileyindia](http://google.com/+wileyindia)

SHELVING CATEGORY  
Management

## Wiley India Pvt. Ltd.

4435-36/7, Ansari Road, Daryaganj  
New Delhi-110 002  
Customer Care +91 11 43630000  
Fax +91 11 23275895  
[csupport@wiley.com](mailto:csupport@wiley.com)  
[www.wileyindia.com](http://www.wileyindia.com)

WILEY

ISBN 978-81-265-6877-2



9 788126 568772

WILEY